University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2009 *Paper* 254

Readings in Targeted Maximum Likelihood Estimation

Mark J. van der Laan* Sherri Rose†

Susan Gruber‡

*University of California - Berkeley, laan@berkeley.edu

†Division of Biostatistics, University of California, Berkeley, sherrirosephd@gmail.com

‡UC Berkeley, sgruber65@yahoo.com

Readings in Targeted Maximum Likelihood Estimation

Mark J. van der Laan, Sherri Rose, and Susan Gruber

Abstract

This is a compilation of current and past work on targeted maximum likelihood estimation. It features the original targeted maximum likelihood learning paper as well as chapters on super (machine) learning using cross validation, randomized controlled trials, realistic individualized treatment rules in observational studies, biomarker discovery, case-control studies, and time-to-event outcomes with censored data, among others. We hope this collection is helpful to the interested reader and stimulates additional research in this important area.

Readings in Targeted Maximum Likelihood Estimation

First Edition

Edited by:
Mark J. van der Laan
Sherri Rose
Susan Gruber

Contents

Chapter 1

Introduction

We have received many requests for centralized reading material on targeted maximum likelihood estimation. While we are in the process of writing a book on these methods, we decided that it might be helpful to bundle most of our current papers on this topic and post them on http://www.bepress.com/ucbbiostat. In this introductory chapter we present the statistical foundation for targeted maximum likelihood estimation, practical implications of targeted maximum likelihood estimation in randomized controlled trials and observational studies, a comparison to estimating function equation methodology, a methods summary for the applied researcher, and an outline of the papers in this compilation. We hope that *Readings in Targeted Maximum Likelihood Estimation* is helpful to the interested reader and stimulates more research in this important area.

Statistical Foundation for Targeted Maximum Likelihood Estimation

For the sake of context, let's consider the case that one observed n i.i.d. copies of a random variable O with probability distribution P_0, and suppose that one is concerned with estimation and inference for a particular target parameter $\Psi(P_0)$ of this true data generating distribution P_0. Targeted maximum likelihood estimation in semiparametric models for P_0 is the extension of maximum likelihood estimation in parametric models. Three key ingredients are needed for this extension. Firstly, one needs to define the parameter of interest nonparametrically (or semiparametrically) as a function of the data generating distribution varying over the (large) semiparametric model . Many practitioners are used to thinking of their parameter in terms of a regression coefficient, but that luxury is not available in semi or nonparametric models. Instead, one has to carefully think of what feature of the distribution of the data one wishes to target.

Secondly, one needs to estimate the true distribution P_0, or at least, its relevant factor or portion as needed to evaluate the target parameter, and this estimate should respect the actual semiparametric model. As a consequence, nonparametric maximum likelihood estimation is often ill defined or results in a complete overfit, and thereby results in too variable estimators of the target parameter. Therefore, sensible estimation procedures involve putting breaks on algorithms that aim to maximize the log-likelihood (e.g., using greedy algorithms, and a sieve representing a sequence of submodels of the semiparametric model), and then fine

tune the choice of these brakes. We use cross-validation to select these fine tuning parameters. One can come up with a large collection of equally appropriate algorithms and fine tuning parameters, resulting in a library of candidate estimators of the distribution of the data. Our research papers on cross-validation, starting in 2003 (*Unified Cross-Validation Methodology For Selection Among Estimators and a General Cross-Validated Adaptive Epsilon-Net Estimator: Finite Sample Oracle Inequalities and Examples*), have focused on understanding the properties of the cross-validation selector for any type of loss function, including the log-likelihood loss function. The theoretical results obtained for the cross-validation selector in this paper have inspired us to propose a general super learning methodology for estimation of distributions of the data, or factors of the distributions of the data. This super learning methodology takes as input a library of candidate estimators of the distribution of the data, and then uses cross-validation to determine the best weighted combination of these estimators. It is assumed or arranged that the loss function is uniformly bounded so that our oracle results for the cross-validation selector apply. The super learning methodology results now in an estimator of the distribution of the data that will be inputted as an initial estimator in the targeted maximum likelihood procedure. This initial estimator is optimized with respect to (w.r.t.) a global loss function such as the log-likelihood loss function, and is thereby not targeted towards the target parameter, ψ_0. That is, it will be too biased for ψ_0 due to a bias variance trade-off w.r.t to the more ambitious full P_0 instead of having used a bias-variance trade-off w.r.t ψ_0. The targeted maximum likelihood step is tailored to remove bias due to the non-targeting.

The targeted maximum likelihood step involves now updating of this initial (super learning based) estimator of P_0 to tailor its fit to estimation of the target $\Psi(P_0)$. This is carried out by determining a fluctuation function applied to the initial estimator with a fluctuation parameter ϵ, where fitting ϵ is the (asymptotic) equivalent of fitting $\Psi(P_0)$ in the semi-parametric model. One now estimates ϵ with maximum likelihood estimation (like maximum likelihood estimation in a parametric model), and updates the initial estimator accordingly. If needed, this updating step is iterated till convergence, and the final update $\hat{P}^*$ is called the targeted maximum likelihood estimator of P_0, while the resulting substitution estimator $\Psi(\hat{P}^*)$ of $\Psi(P_0)$ is the targeted maximum likelihood estimator of ψ_0. This targeted maximum likelihood step uses maximum likelihood fitting of the data to obtain a bias reduction for the target $\Psi(P_0)$.

An important feature of the targeted maximum likelihood estimator is that it solves the efficient influence curve/score equation: if $D^*(P)$ is the efficient influence curve at P, and $\hat{P}^*$ is the targeted maximum likelihood estimator of P_0, then

$$0 = \sum_{i=1}^{n} D^*(\hat{P}^*)(O_i).$$

This can then be used to establish that targeted maximum likelihood estimator is asymptotically efficient if the initial estimator is consistent, and remarkably robust in the sense that for many data structures and semiparametric models, the targeted maximum likelihood estimator of ψ_0 remains consistent even if the initial estimator is inconsistent. In particular, in censored data and causal inference models, the targeted maximum likelihood estimator is a so called double robust estimator: in such semiparametric models the density $d\hat{P}^*$ of targeted maximum likelihood estimator $\hat{P}^*$ can be factorized as $d\hat{P}^* = \hat{Q}^*\hat{g}^*$, where $\hat{g}^*$ is the

estimator of the censoring and treatment mechanism, and the targeted maximum likelihood estimator $\Psi(\hat{Q}^*)$ of $\psi_0 = \Psi(Q_0)$ is consistent if either $\hat{Q}^*$ or $\hat{g}^*$ is consistent.

Practical Implications of Targeted Maximum Likelihood Estimation

The double robustness of the targeted maximum likelihood estimator has important implications for both the analysis of randomized clinical trials as well as observational studies. In a randomized clinical trial (RCT) the treatment assignment process is known, and it is often assumed that missingness or drop-out is non-informative. When this assumption holds, the $\hat{g}$, comprising the treatment and censoring mechanism, is always correctly estimated, and therefore the targeted maximum likelihood estimator will provide valid type-I error control and confidence intervals for the causal effect of the investigated treatment. Moreover, the use of targeted maximum likelihood estimation often results in efciency gains with respect to the unadjusted estimator commonly employed in the analysis of RCT data. There are two reasons for this. First, the unadjusted estimator is restricted to considering only complete cases, ignoring observations where the outcome is missing. The targeted maximum likelihood approach integrates over all observations. Second, targeted maximum likelihood estimation can exploit information in measured baseline and time-dependent covariates. This allows for bias reduction due to empirical confounding. Perhaps more importantly, it naturally adjusts for drop-out/missingness as well, and can also be used to assess the estimate of the effect of treatment under non-compliance. Unlike an unadjusted estimator, targeted maximum likelihood estimation does not rely on an assumption of non-informative missing/drop-out. Pre- specication of the targeted maximum likelihood estimator in the statistical analysis plan allows for appropriate adjustment by measured confounders while avoiding the possible introduction of bias should that decision be based on human intervention. Therefore, targeted maximum likelihood estimators can be used for both the efcacy as well as the safety analysis in Phase II, III, IV clinical trials.

As a simple example of the potential gain in efficiency obtained with targeted maximum likeihood estimation, the relative efficiency of the targeted maximum likelihood estimator relative to the unadjusted estimator of the causal additive risk in a standard randomized control trial with two arms, no missingness or censoring, is given by 1 minus the R-square of the regression of the clinical outcome Y on the baseline covariates W implied by the targeted maximum likelihood fit of the regression of Y on the binary treatment and baseline covariates. That is, if the baseline covariates are predictive, one will gain efficiency, and one can predict the amount of improvement from the actual regression fit. This does not take into account the additional savings obtained by the bias reduction of the targeted maximum likelihood estimator relative to the unadjusted estimator. That is, in randomized controlled trials, including sequentially randomized controlled trials, one can still fully respect the likelihood of the data and obtain fully efficient and unbiased estimators, without taking the risk of bias due to model misspecification (which has been the sole reason for the application of inefficient unadjusted estimators). On the contrary, the better one fits the models, as can be evaluated with the cross-validated log-likelihood, the more bias reduction and efficiency gain will have been achieved.

In both randomized trials and observational studies, the utilization of efficient and maximally unbiased estimators is extremely important. One cannot analyze the effect of high dose of a drug on heart attack in a post-market safety analysis using parametric logistic regression or Cox-proportional hazards models, and put much trust in a p-value. It is already a priori known that these models are biased and that the effect estimate will be estimating this bias, so that under the null hypothesis of no treatment effect, the resulting test statistic will reject the null hypothesis wrongly with probability tending to 1 as sample size increases.

As a consequence, the only alternative is to use semiparametric models that acknowledge what is known and what is not known, and use robust and efficient estimators in a semiparametric model. Given such infinite dimensional semiparametric models, we need to employ machine learning, and, in fact, as theory suggests, we should not be married to one particular machine learning algorithm, but let the data speak by using super learning. That is, one cannot foresee what kind of algorithm should be used, but one should build a rich library of approaches, and use cross-validation to combine these estimators into an improved estimator that adapts the choice to the truth. In addition, again, as theory teaches us, we have to target the fit towards the parameter of interest, to remove bias for the target parameter, and to improve the statistical inference based on the central limit theorem. Targeted maximum likelihood estimation combined with super learning provides such an approach, while we maintain the log-likelihood as the principle criterion.

Targeted maximum likelihood estimation distinguishes from estimating equation methodology (e.g., see the book **Unified Methods for Censored Longitudinal Data and Causality**, van der Laan and Robins, 2003) and (regularized) maximum likelihood estimation, but it also inherits the good properties of both. Targeted maximum likelihood estimation distinguishes from nonparametric or regularized maximum likelihood estimation by fully utilizing the power of cross-validation (super learning) to fine-tune the bias-variance trade-off w.r.t. the distribution P_0 of the data , thereby increasing adaptivity to the true P_0, and by targeting the fit to remove bias w.r.t. ψ_0. In particular, it achieves higher rates of convergence for P_0 itself, higher efficiency due to better fit of true P_0, or even higher rates of convergence for ψ_0, it is less biased for ψ_0 due to the targeted maximum likelihood step, and, as a bonus, the statistical inference based on the central limit theorem is also heavily improved relative to just using a regularized maximum likelihood estimator.

Just as an example illustrating that a regularized maximum likelihood estimator is not targeted towards the target, a typical machine learning algorithm for prediction might not select the treatment variable so that the resulting treatment effect or variable importance equals zero. Such an estimate is not helpful, and follows a heavily non-normal distribution (it will have a pointmass at zero). Similarly, a kernel density estimator with an optimally selected bandwidth (e.g., based on likelihood based cross-validation) will result in a survival function with a bias that converges to zero at a slower rate than $1/\sqrt{n}$ (n is sample size), so that the substitution estimator of a survival function at a point based on this optimal kernel density estimator will have an asymptotic relative efficiency of zero (!) relative to the simple empirical survival function. However, if we apply the targeted maximum likelihood estimation step to the kernel density estimator, then the resulting targeted maximum likelihood estimator of the survival function is efficient, and it would also have been efficient if the kernel density estimator would be replaced by a wrong guess of the true density. The point is: the best estimator of a density is not a good enough estimator of a smooth feature

of the density, but the targeted maximum likelihood estimation step takes care of this.

Advantages over Estimating Equation Methods

In comparison with locally efficient estimating equation methodology (e.g., augmented IPCW-estimator in causal inference and censored data models) the locally efficient targeted maximum likelihood estimation, has the following advantages:

No need for estimating function: The estimating equation methodology relies on representing the efficient score/influence curve $D^*(P)$, the so called canonical gradient of the pathwise derivative of the parameter Ψ at P, as an estimating function in the parameter ψ of interest, and nuisance parameters: $D^*(P) = D(\Psi(P), \eta(P))$. This restricts the estimating equation methodology to parameters for which such a representation of an efficient influence curve $D^*(P) = D(\Psi(P), \eta(P))$ is possible. This is an important and unnecessary restriction.

Targeted maximum likelihood estimation uses the efficient influence curve $D^*(P)$ at P to define the fluctuation function applied to an initial P. This does not require that the efficient influence curve, D^*, also be an estimating function. Therefore targeted maximum likelihood estimation can still be used in situations where estimating equation methodology cannot be applied due to the efficient influence curve not being an estimating function in the parameter of interest, such as, for example, when the parameter of interest is a nonparametric extension of the log-rank parameter.

Respects global constraints of model: The targeted maximum likelihood estimator of ψ_0 is obtained by substitution of an estimator $\hat{P}^*$ in the model into the parameter mapping $\Psi()$. As a consequence, it respects the knowledge of the model.

On the other hand, an estimator of ψ_0 that is obtained as a solution of an estimating equation such as $0 = \sum_i D^*(\psi, \hat{\eta})(O_i)$ is often *not* a substitution estimator: i.e., it cannot be written as $\Psi(\hat{P})$ for a specified estimator $\hat{P}$ in the model. To be specific, suppose one wishes to estimate the treatment specific mean $EY(1) = E_W E(Y \mid A = 1, W)$ based on n i.i.d. copies of (W, A, Y), Y being binary. Then the estimator ψ_n solving the efficient influence curve estimating equation (i.e., the augmented IPTW-estimator) can fall outside the range $[0, 1]$, due to inverse probability of treatments being close to zero. This results in a loss of efficiency and truncation of the estimate has its own obvious problems. On the other hand, the targeted MLE of $EY(1)$ will still be between $[0, 1]$.

No need to deal with multiple solutions of the estimating equation: When defining an estimator as a solution of the efficient score/influence curve estimating equation, one often ends up having to solve non-linear equations that can have multiple solutions. The estimating equation itself provides no information on how to select among these candidates for estimation of ψ_0. One can also not use the likelihood since these estimators cannot be represented as $\Psi(\hat{P})$ for some $\hat{P}$, i.e., these are not substitution estimators. This goes back to the basic fact that estimating functions (such as the

efficient score) might not identify the target parameter, and, even if they do, the corresponding estimating equation might not uniquely identify an estimator for a given sample.

Targeted maximum likelihood estimation does not aim to solve an estimating equation and is therefore not affected by this problem.

Log-likelihood of targeted MLE provides direct measure of fit: Consider the example $O = (W, A, Y)$ and $\psi_0 = EY(1)$, as above. Let Q_0 denote the conditional probability distribution of Y, given (A, W), and the marginal probability distribution of W, and let g_0 denote the conditional probability distribution of A, given W. We have $\psi_0 = \Psi(Q_0)$ is only a parameter of this Q_0-factor of the density of P_0. Given an initial estimator $\hat{g}, \hat{Q}$, a targeted maximum likelihood estimator is defined as $\Psi(\hat{Q}^*)$ while an augmented IPTW estimator is defined as the solution in ψ of $0 = \sum_i D^*(\psi, \hat{Q}, \hat{g})(O_i)$. A targeted maximum likelihood can use the log-likelihood fit (i.e., cross-validated) of $\hat{Q}^*$ as a measure of performance of the targeted maximum likelihood estimator of ψ_0. However, the augmented IPTW-estimator cannot be evaluated by the log-likelihood fit of $\hat{Q}$, since $\hat{g}$ is also having an important impact on the estimator. So one might wish to evaluate the performance by evaluating the log-likelihood fit of both $\hat{Q}$ and $\hat{g}$, but the log-likelihood of g is non-informative for parameters of Q_0 due to factorization $dP_0 = Q_0 g_0$ of the density dP_0 of P_0. So one would be using a criterion that is responding to irrelevant features in the data that have nothing to do with estimation of ψ_0. The fact that the estimating equation methodology does not provide a sensible criterion for selecting an estimator of g_0 makes the estimators rely on subjective choices and makes it hard to define a sensible a priori specified estimator.

This happens to be a very helpful advantage of the targeted maximum likelihood estimator. In particular, it allows one to fine tune the $\hat{g}$ (e.g., variable selection, truncation constant) for the sake of applying the fluctuation function to $\hat{Q}$ in the targeted maximum likelihood step, based on the log-likelihood of the corresponding targeted maximum likelihood estimator $\hat{Q}^*$. Due to this feature, we can also fully exploit the oracle properties of the cross-validation selector based on the loss function $-\log Q$ for Q_0 to also make choices about how to estimate g_0 for the sake of making the targeted maximum likelihood step most effective. This inspired the collaborative targeted MLE extension (van der Laan and Gruber, 2009). In particular, it made clear that the estimation of g_0 as required to evaluate the targeted maximum likelihood step should take place in collaboration with the estimation of Q_0.

Methodology Summary

Targeted maximum likelihood estimation is a two-step procedure where one first obtains an estimate of the data-generating distribution P_0. The second stage updates this initial fit in a step targeted towards making an optimal bias-variance trade-off for the parameter of interest $\Psi(P_0)$, instead of the overall density P_0. The procedure is double robust and can incorporate data-adaptive likelihood based estimation procedures to estimate the data-generating distribution and the treatment mechanism. The double robustness of targeted maximum likelihood estimation has important implications in both randomized controlled trials and

observational studies, with potential reductions in bias and gains in efficiency. There are also significant advantages to targeted maximum likelihood estimation methodology over the use of estimating equation methods.

Additionally, we refer readers to Appendix A for an introductory tutorial on targeted maximum likelihood estimation. This Appendix, complete with R code, aims to provide the reader with understanding sufficient to implement a basic version of targeted maximum likelihood estimation. It may be a good starting point for those less familiar with the concepts discussed previously in this introduction.

Outline of the Collection

The papers bundled in *Readings in Targeted Maximum Likelihood Estimation* are some of the fruits of our research over the past years. **Chapter 2** is the original paper *Targeted Maximum Likelihood Learning* (van der Laan and Rubin, 2006) which provides a comprehensive introduction to targeted maximum likelihood estimation, theoretical development, and resulting procedures for the estimation of causal inference, variable importance, and other parameters of interest.

Chapter 3 is titled "Super (Machine) Learning using Cross Validation." It has two parts. The first part, *Super Learner* (van der Laan, et al., 2007), proposes an algorithm for constructing a super learner which uses cross-validation to select weights to combine an initial set of candidate estimators, where the true target (typically a function, such as a conditional density, conditional hazard, regression) is defined as a minimizer of the expectation under the observed data distribution P_0 of a loss function of O and a candidate value for the target. For example, the target could represent the whole distribution of the data, a factor of this distribution, an identifiable part of the distribution of the full underlying data, a regression, a median regression, a conditional hazard, a causal dose response curve, and so on. The second part of Chapter 3, *Loss-Based Cross-Validated Deletion/Substitution/Addition Algorithms in Estimation* (Sinisi and van der Laan, 2004), describes loss-based learning based on cross-validation in the context of regression. This paper discusses the Deletion/Substitution/Addition (DSA) algorithm, which is a data-adaptive model selection procedure based on cross-validation and uses polynomial basis functions to search through a parameter space of potential regression functions. This function is available as an R-package. It illustrates concretely how cross-validaiton is used to make a variety of choices when fitting a regression.

Chapter 4 features *Collaborative Double Robust Targeted Penalized Maximum Likelihood Estimation* (van der Laan and Gruber, 2009). It establishes a new collaborative double robustness result for the targeted maximum likelihood estimator, and, in order to exploit this collaborative robustness, refines the standard targeted maximum likelihood estimation procedure by refining the targeted maximum likelihood step. This involves utilizing likelihood-based cross-validation to select among different targeted maximum likelihood steps possibly indexed by different sets of confounders for the treatment/censoring mechanism, thereby yielding maximally effective bias reduction. We show that if the initial estimator converges fast, then the collaborative targeted maximum likelihood estimator can even be super efficient. It also presents a strategy to penalize the log-likelihood to make the log-likelihood of the targeted maximum likelihood estimation more targeted in the context of sparse data

(i.e., lack of practical identifiability of the target parameter ψ_0), which results in unstable targeted maximum likelihood steps.

Chapter 5, titled "Randomized Controlled Trials," is concerned with targeted maximum likelihood estimation in randomized controlled trials. Firstly, we present *Covariate Adjustment in Randomized Trials with Binary Outcomes: Targeted Maximum Likelihood Estimation* (Moore and van der Laan, 2008), which includes simulation studies assessing potential gains in efficiency one can achieve by having predictive baseline covariates. The targeted maximum likelihood estimator for the data structure $O = (W, A, \Delta, \Delta Y)$ is presented, where W denotes baseline covariates, A treatment, Δ indicator of observing the clinical outcome Y. The paper *Selecting Optimal Treatments Based on Predictive Factors* (Polley and van der Laan, 2009) shows how one can use super learning and targeted maximum likelihood to assess effect modification in clinical trials, one factor at the time, or for estimating the treatment effect as a function of a whole set of baseline covariates. In particular, it allows one to estimate the optimal treatment decision in response to baseline characteristics. The paper *Simple, Efficient Estimators of Treatment Effects in Randomized Trials Using Generalized Linear Models to Leverage Baseline Variables* (Rosenblum and van der Laan, 2009) illustrates that the results for the targeted maximum likelihood estimator prove that misspecified generalized linear regression models provide valid estimates of marginal causal effects in randomized controlled trials. That is, these misspecified regression estimators represent particular implementations of the targeted maximum likelihood estimator in randomized controlled trials, and are thereby guaranteed to be consistent for the target.

Chapter 6 features *Estimating the Effect of Vigorous Physical Activity on Mortality in the Elderly Based on Realistic Individualized Treatment and Intention-to-Treat Rules* (Bembom and van der Laan, 2007), which presents a practical illustration of the importance of realistic individualized treatment rules in causal inference. It applies targeted maximum likelihood estimation to estimate these causal effects defined by realistic treatment rules in populations where certain levels of treatment are unlikely to be observed in some individuals. Since this is the first application of targeted maximum likelihood estimation to estimate the effect of individualized treatment rules we included this paper in this particular collection of readings, although, we have earlier work (van der Laan and Petersen, 2006, among others) on realistic rules for multiple time-point treatment interventions, but these previous papers apply the IPCW-estimator.

Chapter 7, "Biomarker Discovery," concerns the application of targeted maximum likelihood estimation in biomarker discovery. The paper *Targeted Methods for Biomarker Discovery, the Search for a Standard* (Tuglus, van der Laan, 2008) proposes targeted maximum likelihood estimators of variable importance (tVIM) as a standardized method for biomarker discovery. In this paper we focus on variable importance analysis of possibly continuous variables, exploiting the semiparametric regression model to define variable importance. Simulations and data analyses are used to illustrate the benefits achieved in biomarker discovery relative to current approaches for variable importance analyses (univariate regression, random forest, lars). The paper *Biomarker Discovery using Targeted Maximum Likelihood Estimation: Application to the Treatment of Antiretroviral Resistant HIV Infection* (Bembom, et al., 2008) discusses and implements targeted maximum likelihood estimation for variable importance for a set of candidate binary biomarkers such as mutations or single

nucleotide polymorphisms. The paper *Data-adaptive Selection Of The Adjustment Set In Variable Importance Estimation* (Bembom, et al., 2008) introduces an algorithm intended to make variable importance estimation more robust with respect to violations of the experimental treatment assignment assumption. This algorithm is applied to a dataset in an effort to identify mutations in the protease enzyme of HIV that have an effect on virologic response to the commonly used antiretroviral drug lopinavir.

Chapter 8, "Case-Control Studies," presents targeted maximum likelihood estimation, and, in particular, targeted likelihood based causal inference for case-control studies. The paper *Estimation Based on Case-Control Designs with Known Prevalance Probability* (van der Laan, 2008) provides a comprehensive introduction to case-control weighted targeted maximum likelihood estimation theory for case-control study designs. *Simple Optimal Weighting of Cases and Controls in Case-Control Studies* (Rose and van der Laan, 2008) implements case-control weighted targeted maximum likelihood estimation for independent case-control study designs, and compares this methodology to existing methods. The paper *Why Match? Investigating Matched Case-Control Study Designs with Causal Effect Estimation* (Rose and van der Laan, 2009) discusses the use of matching in case-control study designs. In particular, it compares the efficiency of matched case-control study designs to independent study designs in varied situations using case-control weighted targeted maximum likelihood estimation. Lastly, *Causal Inference for Nested Case-Control Studies using Targeted Maximum Likelihood Estimation* (Rose and van der Laan, 2009) discusses the use of targeted maximum likelihood estimation in nested case-control study designs. It also compares the efficiency of nested case-control designs to analysis of the full cohort data.

Chapter 9 is titled "Time-to-Event Outcomes and Censored Data." The first paper, *A Note on Targeted Maximum Likelihood and Right Censored Data* (van der Laan and Rubin, 2007), fully develops the targeted maximum likelihood estimator of causal effects in randomized controlled trials with a time-to-event outcome that is subject to right-censoring. The second paper, *Application of Time-to-Event Methods in the Assessment of Safety in Clinical Trials* (Moore, van der Laan, 2009), provides the theoretical ingredients to derive the targeted maximum likelihood estimator for this data structure.

The **Appendix** contains a gentle introduction to and R-code for the targeted maximum likelihood estimator of the causal effect of a binary treatment for the data structure $(W, A, \Delta, \Delta Y)$ allowing for confounding of treatment and missingness of the clinical outcome Y. For the more theoretical oriented reader, the Appendix also includes a collection of worked out examples of targeted maximum likelihood estimation for different data structures and parameters. This shows how the targeted maximum likelihood step is derived, given the data structure and the model. It presents the natural extension of targeted maximum likelihood estimation to targeted minimum loss based learning. In addition, it presents targeted Bayesian learning based on targeted maximum likelihood, presenting a mapping from a prior distribution on the target parameter into a targeted (bias reduced) posterior distribution of the target parameter.

For the interested reader, we note that targeted maximum likelihood has been generalized to group sequential adaptive designs in which the censoring and treatment mechanism of a new subject/unit can be adapted in response to the observed data on the previously recruited units: *The Construction and Analysis of Adaptive Group Sequential Designs* (van der Laan, 2008). The latter paper contains many additional examples of targeted maximum

likelihood estimation for longitudinal data structures, including effects of multiple time-point interventions, and shows that adaptive designs can learn the optimal design in a group sequential design, while preserving frequentist statistical inference based on the martingale central limit theorem.

Finally, we remark that Target Analytics, Inc. (www.targetanalytics.com) is a company founded on the premise of implementing statistical software based on targeted maximum likelihood estimation. A version of TargetDiscovery, a variable importance software product based on targeted maximum likelihood estimation of variable importance across a user supplied set of target variables, can be tested directly on the website. The IP for the targeted maximum likelihood estimation methodology is owned by University of California, Berkeley.

Chapter 2

Targeted Maximum Likelihood Estimation

2.1 *Targeted Maximum Likelihood Learning*

The following article appears as it was published in the *International Journal of Biostatistics* in 2006, http://www.bepress.com/ijb/vol2/iss1/11/.

It was originally published on the University of California, Berkeley Division of Biostatistics Working Paper Series website in 2006, http://www.bepress.com/ucbbiostat/paper213/.

Targeted Maximum Likelihood Learning

Mark J. van der Laan and Daniel B. Rubin

Suppose one observes a sample of independent and identically distributed observations from a particular data generating distribution. Suppose that one has available an estimate of the density of the data generating distribution such as a maximum likelihood estimator according to a given or data adaptively selected model. Suppose that one is concerned with estimation of a particular pathwise differentiable Euclidean parameter. A substitution estimator evaluating the parameter of the density estimator is typically too biased and might not even converge at the parametric rate: that is, the density estimator was targeted to be a good estimator of the density and might therefore result in a poor estimator of a particular smooth functional of the density. In this article we propose a one step (and, by iteration, k-th step) targeted maximum likelihood density estimator which involves 1) creating a hardest parametric submodel with parameter epsilon through the given density estimator with score equal to the efficient influence curve of the pathwise differentiable parameter at the density estimator, 2) estimating epsilon with the maximum likelihood estimator, and 3) defining a new density estimator as the corresponding update of the original density estimator. We show that iteration of this algorithm results in a targeted maximum likelihood density estimator which solves the efficient influence curve estimating equation and thereby yields a locally efficient estimator of the parameter of interest, under regularity conditions. In particular, we show that, if the parameter is linear and the model is convex, then the targeted maximum likelihood estimator is often achieved in the first step, and it results in a locally efficient estimator at an arbitrary (e.g., heavily misspecified) starting density. This tool provides us with a new class of targeted likelihood based estimators of pathwise differentiable parameters. We also show that the targeted maximum likelihood estimators are now in full agreement with the locally efficient estimating function methodology as presented in Robins and Rotnitzky (1992) and van der Laan and Robins (2003), creating, in particular, algebraic equivalence between the double robust locally efficient estimators using the targeted maximum likelihood estimators as an estimate of its nuisance parameters, and

targeted maximum likelihood estimators. In addition, it is argued that the targeted MLE has various advantages relative to the current estimating function based approach. We proceed by providing data driven methodologies to select the initial density estimator for the targeted MLE, thereby providing data adaptive targeted maximum likelihood estimation methodology. Finally, in our accompanying technical report we show that targeted maximum likelihood estimation can be generalized to estimate any kind of parameter, such as infinite dimensional non-pathwise differentiable parameters, by restricting the likelihood and cross-validated log-likelihood to targeted candidate density estimators only. We illustrate the method with various worked out examples.

1 Introduction

Let $O_1, \ldots, O_n$ be n independent and identically distributed (i.i.d.) observations of an experimental unit O with probability distribution $P_0 \in \mathcal{M}$, where $\mathcal{M}$ is the statistical model. For the sake of presentation, we will assume that $\mathcal{M}$ is dominated by a common measure μ so that we can identify each possible probability measure $P \in \mathcal{M}$ by its density $p = dP/d\mu$. In the discussion we point out that our methods are not restricted to models dominated by a single measure. Let P_n be the empirical probability distribution of $O_1, \ldots, O_n$ which puts mass $1/n$ on each of the n observations. Let $p_0 = \frac{dP_0}{d\mu}$ be the density of p_0 with respect to a dominating measure μ, and let p_n be a density estimator of p_0. For example, $p_n \equiv \Phi(P_n)$ could be the maximum likelihood estimator defined by the following mapping Φ

$$p_n = \Phi(P_n) \equiv \arg\max_{P \in \mathcal{M}} \sum_{i=1}^{n} \log \frac{dP}{d\mu}(O_i).$$

Alternatively, if the model $\mathcal{M}$ is too large in the sense that the maximum likelihood estimator is too variable or even inconsistent, then one typically proposes a sieve $\mathcal{M}_s \subset \mathcal{M}$, indexed by indices s, approximating $\mathcal{M}$, and computes candidate maximum likelihood estimators

$$p_{ns} = \Phi_s(P_n) \equiv \arg\max_{P \in \mathcal{M}_s} \sum_{i=1}^{n} \log \frac{dP}{d\mu}(O_i).$$

In such a setting it remains to data adaptively select s. For example, one could use likelihood based cross-validation to select s:

$$s_n = \arg\max_{s} E_{B_n} \sum_{i:B_n(i)=1} \log \Phi_s(P^0_{n,B_n})(O_i),$$

where $B_n \in \{0,1\}^n$ is a random vector of binary variables defining a random split in a training sample $\{i : B_n(i) = 0\}$ and validation sample $\{i : B_n(i) = 1\}$, and P^0_{n,B_n}, P^1_{n,B_n} denote the empirical probability distributions of the training and validation sample, respectively. Now, one would define the estimator of p_0 as the cross-validated maximum likelihood estimator given by

$$p_n = \Phi(P_n) \equiv p_{ns_n} = \Phi_{s_n}(P_n).$$

It is common practice to evaluate one or many Euclidean valued smooth functionals $\Psi(p_n)$ of the density estimator p_n and view them as estimators of the parameter $\Psi(p_0)$ for given parameter mappings $\Psi : \mathcal{M} \to \mathbb{R}^d$. Although

this method is known to result in efficient estimators of $\Psi(p_0)$ in parametric models (i.e., $\mathcal{M}$ in the above definition of p_n is a parametric model), in general, such substitution estimators are not correctly trading off bias and variance with respect to the parameter of interest $\psi_0 = \Psi(p_0)$. For example, a univariate (standard) kernel density estimator optimizing the mean squared error with respect to p_0, assuming a continuous second derivative, can have bias of the order $n^{-2/5}$ based on an optimal bandwidth of the order $n^{-1/5}$. The corresponding substitution estimator of the cumulative distribution function at a point can have bias which converges to zero at the same rate $n^{-2/5}$, but a variance of $O(1/n)$, so that the substitution estimator has a variance $(1/n)$ which is smaller than the square bias $(n^{-4/5})$ by an order of magnitude. In particular, the smoothed empirical cumulative distribution functions would not even converge at root-n rate due to the fact that $\sqrt{n}$ times the bias $n^{-2/5}$ does not converge to zero: that is, in this kernel density estimator example $\sqrt{n}n^{-2.5} \to \infty$, so that the relative efficiency of the empirical cumulative distribution function and this smooth cumulative distribution function converges to zero. This shows that substitution estimators based on optimal (*for the purpose of the density itself*) density estimators of the cumulative distribution function are typically theoretically inferior to other more targeted estimators of the parameter of interest. In general, substitution estimators based on density estimators might simply not be very good estimators, and, in particular, likelihood based substitution estimators will often fail to be asymptotically efficient due to the bias caused by the curse of dimensionality: the kernel density example already shows the failure of likelihood based learning of smooth parameters of a density of a univariate random variable, and it gets much worse for densities of multivariate random variables. This issue has been stressed repeatly by Robins and co-authors (see e.g., Robins and Rotnitzky (1992) and van der Laan and Robins (2003)). This article proposes a method which, given a particular pathwise differentiable parameter of interest, allows one to map a density estimator (such as p_n or p_{ns} for each s) into a targeted maximum likelihood density estimator so that the corresponding substitution estimator of ψ_0 is locally efficient, under reasonable conditions: that is, if the starting density estimator is consistent, it will typically be efficient, and otherwise in certain classes of problems it might still be consistent and asymptotically linear.

Specifically, in this article we propose a one step maximum likelihood density estimator which involves 1) creating a parametric model with Euclidean parameter ϵ (e.g., the same dimension d as the parameter ψ_0) through a given density estimator p_n^0 (e.g., s-specific MLE p_{ns}) at $\epsilon = 0$ whose scores include the components of the efficient influence curve of the pathwise differentiable parameter at the density estimator p_n^0, 2) estimating ϵ with the maximum

likelihood estimator of this parametric model, and 3) defining a new density estimator p_n^1 as the corresponding fluctuation of the original density estimator p_n^0. In addition, iterating this process results in a sequence of p_n^k with increasing log-likelihood converging to a solution of the efficient influence curve estimating equation, and thereby typically results in a locally efficient substitution estimator of ψ_0. We refer to this solution as the targeted maximum likelihood estimator based on the initial p_n^0. We provide various examples in which this targeted maximum likelihood estimator is achieved at the first step of the algorithm.

In particular, one can map each model based MLE p_{ns} into a targeted MLE p_{ns}^* (targeted towards ψ_0). We suggest that it is appropriate to select among this collection of targeted MLEs p_{ns}^* with likelihood based cross-validation, as explained heuristically in our accompanying technical report: targeted MLE's are comparable w.r.t. to being fully trained w.r.t. estimation of the parameter of interest, which makes the log-likelihood an appropriate criteria to select among them. That is, let $p_{ns}^* = \hat{\Phi}_s^*(P_n)$ be the s-specific targeted MLE applied to the initial density estimator p_{ns}. Let

$$s_n = \arg\max_s E_{B_n} \sum_{i:B_n(i)=1} \log \hat{\Phi}_s^*(P_{n,B_n}^0)(O_i),$$

where $B_n \in \{0,1\}^n$ is a random vector of binary variables defining a random split in a training sample $\{i : B_n(i) = 0\}$ and validation sample $\{i : B_n(i) = 1\}$, and P_{n,B_n}^0, P_{n,B_n}^1 denote the empirical probability distributions of the training and validation sample, respectively, as above. Now, likelihood cross-validated targeted MLE is defined as:

$$p_n^* = \hat{\Phi}(P_n) \equiv p_{ns_n}^* = \hat{\Phi}_{s_n}^*(P_n).$$

We also note that the candidate models indexed by s can be chosen to represent a sieve in a possibly misspecified (big) model $\mathcal{M}$, as long as this model $\mathcal{M}$ is still such that the Kullback-Leibler projection of the true density p_0 on this model identifies the parameter of interest $\Psi(p_0)$ correctly: for example, if the parameter of interest is a parameter of a regression of an outcome Y on covariates W, then one might select as big model the normal densities with unspecified conditional mean, given W, and certain possibly misspecified conditional variance, even though the true density p_0 is not a member of this model.

1.1 Organization of article.

In Section 2, given an initial density estimator p_n^0 (e.g., p_{ns}) of p_0, we formally define the k-th order targeted maximum likelihood density estimator p_n^k, and

corresponding targeted maximum likelihood estimator $\Psi(p_n^k)$ of ψ_0. We illustrate the targeted MLE of the cumulative distribution function at a point in a nonparametric model. In this case, it appears that the first step targeted MLE of ψ_0 algebraically equals the empirical cumulative distribution function, for any given initial density estimator p_n^0. Thus, while the original substitution estimator of the cumulative distribution function would not converge at the parametric rate $1/\sqrt{n}$ due to it being too biased, the first order targeted bias corrected density estimator estimates the cumulative distribution function efficiently. In Section 3 we establish that the targeted MLE solves the efficient influence curve estimating equation, which provides the basis of its asymptotic efficiency for ψ_0. In Section 4 we present general templates for establishing consistency, asymptotic linearity and efficiency of the targeted MLE of ψ_0, which provides a particular powerful theorem for convex models and linear pathwise differentiable parameters stating that the targeted MLE will be consistent and asymptotically linear for an arbitrary starting density, and it will be efficient if the starting (or its targeted MLE version) density consistently estimates the efficient influence curve. We illustrate the latter result with two examples. In Section 5 we discuss the relation, and in particular, the algebraic equivalence, between targeted maximum likelihood estimation and estimating function based estimation if one estimates the nuisance parameters in the estimating functions with the targeted MLE. We point out that targeted MLE is more widely applicable by not relying on being able to map the efficient influence curve in a corresponding estimating function, and it deals naturally with the issue of multiple solutions of estimating equations. In Subsection 5.1 we focus on censored data models to make the comparison with the estimating function methodology in van der Laan and Robins (2003). In particular, we present the targeted MLE approach which results in algebraic equivalence between the Inverse Probability of Censoring Weighted estimator, the double robust IPCW estimator, and the targeted MLE of a parameter of the full data distribution based on observing n i.i.d. observations of a censored data structure under coarsening at random (CAR). These results show that the targeted MLE does not only provide a boost for likelihood based estimation, but it also provides an improvement relative to the current implementation of locally efficient estimation based on estimating function methodology. In Section 6 we present important examples illustrating the power and computational simplicity of this new targeted maximum likelihood estimator: estimation of a marginal causal effect, and the parametric component in a semiparametric regression model, and we present a simulation to illustrate the targeted MLE. In Section 7 we present a loss based approach of targeted MLE learning based on the unified loss function based approach in van der Laan and Dudoit

(2003). We end this article with a discussion in Section 8. In our accompanying technical report we show generalizations of the targeted MLE of pathwise differentiable parameters to targeted MLE of general parameters.

1.2 Some relevant literature overview.

There exist various methods for construction of an efficient estimator of a parameter based on parametric models. In particular, Fisher's method of maximum likelihood estimation can be applied, or closely related M-estimate (i.e., estimators defined as solutions of estimating equations) methods which work under minimal conditions. Maximum likelihood estimation in semiparametric models has been an extensive research area of interest. Here we suffice with a referral to van der Vaart and Wellner (1996b) for a partial overview of the theory for the analysis of maximum likelihood. There are plenty of examples in which the straightforward semiparametric MLE even fails to be consistent, but often an appropriate regularization can be applied to repair the consistency of the semiparametric MLE: e.g., see van der Laan (1995) for such examples based on censored data. However, as argued above in the kernel density estimator example, maximum likelihood based smoothing/model selection will often provide the wrong trade-off of bias and variance for specific smooth parameters. The literature (notably Robins and co-authors) has recognized this problem with likelihood based estimation. For example, smoothing survival functions or smoothing the nonparametric components in a semiparametric regression model requires so called "under-smoothing" in order to obtain root-n consistency for the parameter of interest: see e.g., Cosslett (2004).

For an overview of the literature on efficient estimation of pathwise differentiable parameters in semiparametric models we refer to Bickel et al. (1993b). In particular, the latter presents the general one step estimator based on an estimate of the efficient influence curve: see e.g. Klaassen (1987). For an overview of the literature on locally efficient estimating function based estimation of pathwise differentiable parameters based on censored longitudinal data (starting with the ground breaking paper Robins and Rotnitzky (1992)), we refer to van der Laan and Robins (2003).

A unified loss function approach based methodology for estimation and estimator selection, and concrete illustration of this method in various examples is presented in van der Laan and Dudoit (2003). This methodology is general by allowing the loss function to be an unknown function of the experimental unit and the parameter values. van der Laan and Rubin (2005) and van der Laan and Rubin (2006) present an alternative unified estimating function methodology for both estimation and estimator selection. The latter

two methodologies provide two general strategies for data adaptive estimation of any parameter in any model.

We note that these (unified) loss function and (unified) estimating function based approaches give up on using the log-likelihood as loss function for the purpose of estimator selection and estimation when the parameter of interest is not the actual density of the data, but a particular parameter of it: these methods replace the log-likelihood loss function by a loss function or an estimating function targeted at the parameter of interest. From that point of view, the current article shows that it is not necessary to replace the log-likelihood loss function by a targeted loss function, but that one can also target the directions in which one maximizes the log-likelihood.

2 Targeted maximum likelihood estimators.

Let $\Psi : \mathcal{M} \to \mathbb{R}^d$ be a pathwise differentiable parameter at any density $p \in \mathcal{M}$, where $\mathcal{M}$ denotes the statistical model consisting of the possible densities $p = dP/d\mu$ of O with respect to some dominating measure μ. That is, given a sufficiently rich class of one-dimensional regular parametric submodels $\{p_\delta : \delta\}$ with parameter δ of $\mathcal{M}$ through the density p at $\delta = 0$, we have for each of these submodels p_δ with score s at $\delta = 0$ and $p_{\delta=0} = p$

$$\frac{d}{d\delta} \Psi(p_\delta)|_{\delta=0} = E_p S(p)(O)s(O)$$

for some $S(p) \in (L_0^2(p))^d$, where $L_0^2(p)$ denotes the Hilbert space of functions of O with mean 0 and finite variance under P, endowed with inner product $\langle h_1, h_2 \rangle_P = E_p h_1(O)h_2(O)$. This random variable $S(p) \in (L_0^2(p))^d$ is called a gradient of the pathwise derivative at p. Let $T(p) \subset L_0^2(p)$ be the tangent space at p which is defined as the closure of the linear span of the scores s of this class of submodels through p. If the model is not locally saturated in the sense that $T(p) = L_0^2(p)$, then there can be many gradients. Let $T_{nuis}^{\perp}(p) \subset L_0^2(p)$ be the orthogonal complement of the so called nuisance tangent space, where the latter is defined as the closure of the linear span of all scores of p_δ for which the pathwise derivative equals 0 (see van der Laan and Robins (2003), Chapter 1). As in van der Laan and Robins (2003), we denote the set of gradients at p with $T_{nuis}^{\perp *}(p) \subset (T_{nuis}^{\perp}(p))^d$. Let $S^*(p)$ be the so called canonical gradient which is the unique gradient whose d components $S^*(p)_j$, $j = 1, \ldots, d$, are elements of the tangent space $T(P)$. A submodel $\{p_\epsilon : \epsilon\}$ with score $S^*(p)$ at $\epsilon = 0$ is often referred to as a hardest submodel (Bickel et al. (1993a)), as we will also do in this article.

Let $(O,p) \rightarrow D(p)(O)$ be a point-wise well defined class of functions on the Cartesian product of the support of O and the model $\mathcal{M}$, which satisfies

$$D(p) = S^*(p) \ P_0\text{-a.e. for all } p \in \mathcal{M}.$$

As an example, consider letting O be a Euclidean valued d-variate random variable with density p_0. Let $\mathcal{M}$ be the class of all continuous densities with respect to Lebesgue measure μ, and let $\Psi(p) = \int_0^t p(o)d\mu(o)$ be the cumulative distribution function at a point $t \in \mathbb{R}$ corresponding with density p. In this case $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ is pathwise differentiable parameter at p with efficient influence curve $S(p)(O) = I(O \leq t) - \Psi(p)$, and, because the model is locally saturated, it is also the only influence curve/gradient. So $D(p) = I(O \leq t) - \Psi(p)$. Similarly, given a set of user supplied points $\{t_1, \ldots, t_d\}$, we could define the d-dimensional Euclidean parameter $\Psi(p) = (\Psi(p)(t_j) \equiv \int_0^{t_j} p(o)d\mu(o) : j = 1, \ldots, d)$ representing the cumulative distribution function at d points. In this case, $D(p) = (I(O \leq t_j) - \Psi(p)(t_j) : j = 1, \ldots, d)$ has d components.

A general methodology for construction of functions $D_h(p)$ indexed by an $h \in \mathcal{H}$ so that $\{D_h(p) : h \in \mathcal{H}\} \subset T^{\perp}_{nuis}(p)$ (or equality) is presented in van der Laan and Robins (2003). In van der Laan and Robins (2003) the class of functions $\{D_h(p) : h \in \mathcal{H}\}$ is referred to as a representation of the orthogonal complement of the nuisance tangent space, which is then used to map into a class of corresponding estimating functions for the pathwise differentiable parameter $p \rightarrow \Psi(p)$ of the form $p \rightarrow D_h(\Psi(p), \Upsilon(p))$ with Υ representing a nuisance parameter. In van der Laan and Robins (2003), for a variety of general classes of models and censored data structures O, explicit representations of the orthogonal complement of the nuisance tangent space, $T^{\perp}_{nuis}(p)$, corresponding gradients, $T^{\perp *}_{nuis}(p)$, and canonical gradient $S^*(p)$, have been provided.

Let $p_n^0 = \Phi(P_n) \in \mathcal{M}$ be a density estimator of $p_0 = dP_0/d\mu$. Define now a parametric submodel $\{p_n^0(\epsilon) : \epsilon \in \mathbb{R}^k\} \subset \mathcal{M}$ through p_n^0 at $\epsilon = 0$ whose linear span of scores of ϵ at $\epsilon = 0$ includes all d components of $D(p_n)$. One possibility is to choose $\epsilon \in \mathbb{R}^d$ of the same dimension as $D(p)$ and arrange that the score of ϵ_j at $\epsilon = 0$ equals $D_j(p)$, $j = 1, \ldots, d$. For example, if the model $\mathcal{M}$ is convex then the following model typically applies

$$p_n^0(\epsilon) \equiv (1 + \epsilon^{\top} D(p_n^0))p_n^0, \tag{1}$$

where $\epsilon \in \mathbb{R}^d$ denotes the parameter ranging over all values for which $p_n^0(\epsilon)$ is a proper density. Note that indeed $p_n^0(0) = p_n^0$, $p_n^0(\epsilon)$ is a density (positive valued and integrates till 1) for ϵ small enough, and $\frac{d}{d\epsilon} \log p_n^0(\epsilon)\Big|_{\epsilon=0} = D(p_n^0)$.

One can also use an exponential family

$$p_n^0(\epsilon) \equiv C(\epsilon, p_n^0) \exp(\epsilon^\top D(p_n^0)) p_n^0$$

for $C(\epsilon, p_n^0)$ be a normalizing constant. In general, one can choose a parameterization $\epsilon \to p_n^0(\epsilon) \in \mathcal{M}$ which is smooth in ϵ at $\epsilon = 0$ and whose score at $\epsilon = 0$ equals $D(p_n^0)$. However, we will also consider submodels $p_n^0(\epsilon)$ with additional scores in order to arrange that the targeted MLE will be fully targeted towards estimation of $D(p_0)$.

Let

$$\epsilon_n = \epsilon(P_n \mid p_n^0) \equiv \arg \max_{\{\epsilon : p_n^0(\epsilon) \in \mathcal{M}\}} \sum_{i=1}^n \log p_n^0(\epsilon)(O_i)$$

be the maximum likelihood estimator of ϵ treating the density estimator p_n^0 as given and fixed. We will assume that the maximum is attained in the interior of $\mathcal{M}$ so that ϵ_n solves the estimating equation:

$$0 = P_n \frac{\frac{d}{d\epsilon} p_n^0(\epsilon)}{p_n^0(\epsilon)}.$$

Here we use the common notation $Pf \equiv \int f(o) dP(o)$. For example, if $p_n^0(\epsilon) = (1 + \epsilon^\top D(p_n^0)) p_n^0$, as one might choose in convex models, then we have that ϵ_n is the solution of

$$0 = \frac{1}{n} \sum_{i=1}^n \frac{D(p_n^0)(O_i)}{1 + \epsilon_n^\top D(p_n^0)(O_i)}.$$

This defines now an updated density estimator

$$p_n^1 \equiv p_n^0(\epsilon_n) = p_n^0(\epsilon(P_n \mid p_n^0)) \in \mathcal{M}.$$

Note that this simply defines a method for mapping an initial density estimator $p_n^0 \in \mathcal{M}$ in a new density estimator $p_n^1 \in \mathcal{M}$, which we call the first step targeted maximum likelihood estimator. By iterating this process one obtains the k-step targeted maximum likelihood estimator p_n^k, $k = 1, \ldots$

Definition 1 *Given an initial density estimator $p_n^0 = \hat{\Phi}^0(P_n)$ based on the empirical probability distribution P_n, a parametric fluctuation $\{p_n^0(\epsilon) : \epsilon\} \subset \mathcal{M}$ satisfying $p_n^0(0) = p_n^0$, and $\frac{d}{d\epsilon} \log p_n^0(\epsilon)\Big|_{\epsilon=0} = D^*(p_n^0)$, where the linear span of the components of $D^*(p_n^0)$ include all d components of a canonical gradient $D(p_n^0)$ of the parameter of interest $\Psi : \mathcal{M} \to \mathbb{R}^d$ at p_n^0, a maximum likelihood estimator*

$$\epsilon(P_n \mid p_n^0) \equiv \arg \max_\epsilon \sum_{i=1}^n \log p_n^0(\epsilon)(O_i)$$

of ϵ, we define the first step targeted maximum likelihood density estimator as

$$p_n^1 = \hat{\Phi}^1(P_n) \equiv p_n^0(\epsilon(P_n \mid p_n^0)).$$

This process can be iterated to define the k-step targeted maximum likelihood density estimator as

$$p_n^{k+1} = \hat{\Phi}^{k+1}(P_n) \equiv p_n^k(\epsilon(P_n \mid p_n^k)),\ k = 0, 1, \ldots.$$

The corresponding k-step targeted maximum likelihood estimator of ψ_0 is defined as

$$\hat{\Psi}_k(P_n) = \Psi(p_n^k).$$

The targeted maximum likelihood estimator is defined as

$$\psi_n = \hat{\Phi}^*(P_n) \equiv \lim_{k\to\infty} \Psi(p_n^k),$$

assuming this limit exists.

2.1 Example: Estimating the CDF.

Consider an initial data generating density $p^0 = f$, let $F(t) = \int_{-\infty}^t f(o)do$ denote the associated CDF at some fixed point $t \in \mathbb{R}$, and consider the parametric model

$$\left\{ f_\epsilon(o) = (1 + \epsilon[I(o \le t) - F(t)])f(o) : -\frac{1}{1 - F(t)} \le \epsilon \le \frac{1}{F(t)} \right\}, \quad (2)$$

where one can check that the range restraint on ϵ serves merely to ensure that the family is indeed a proper class of densities. Consider estimating ϵ from maximum likelihood based on an i.i.d. sample $\{O_i\}_{i=1}^n$. The log likelihood is,

$$l(\epsilon) = \sum_{i=1}^n \log(1 + \epsilon[I(O_i \le t) - F(t)]) + \sum_{i=1}^n \log f(O_i). \quad (3)$$

Its derivative is,

$$l'(\epsilon) = \sum_{i=1}^n \frac{I(O_i \le t) - F(t)}{1 + \epsilon[I(O_i \le t) - F(t)]}. \quad (4)$$

Its second derivative is easily seen to be,

$$l''(\epsilon) = -\sum_{i=1}^n \left\{ \frac{I(O_i \le t) - F(t)}{1 + \epsilon[I(O_i \le t) - F(t)]} \right\}^2. \quad (5)$$

Because the log likelihood is concave, we know that the maximum is achieved if $l'(\epsilon) = 0$ has a solution. Letting $F_n(\cdot)$ denote the empirical distribution function, note that we can decompose the terms in $l'(\epsilon)$ into two parts (those for which $I(O_i \leq t)$ are 0 or 1), and the MLE of ϵ can be seen to solve,

$$\begin{aligned} 0 &= l'(\epsilon) \\ &= \sum_{i=1}^{n} \frac{I(O_i \leq t) - F(t)}{1 + \epsilon[I(O_i \leq t) - F(t)]} \\ &= nF_n(t)\frac{1 - F(t)}{1 + \epsilon[1 - F(t)]} + n(1 - F_n(t))\frac{-F(t)}{1 - \epsilon F(t)}. \end{aligned}$$

Moving the second term on the right to the other side of the equation, dividing both sides by n, and multiplying both sides by $(1 + \epsilon[1 - F(t)])(1 - \epsilon F(t))$, the equation reduces to,

$$F_n(t)(1 - F(t))(1 - \epsilon F(t)) = (1 - F_n(t))F(t)(1 + \epsilon(1 - F(t))). \tag{6}$$

This is linear in ϵ, and one can check that the solution is

$$\begin{aligned} \epsilon_n &= \frac{F_n(t)(1 - F(t)) - (1 - F_n(t))F(t)}{F(t)(1 - F(t))} \\ &= \frac{F_n(t) - F_n(t)F(t) - F(t) + F_n(t)F(t)}{F(t)(1 - F(t))} \\ &= \frac{F_n(t) - F(t)}{F(t)(1 - F(t))}. \end{aligned} \tag{7}$$

Because $0 \leq F_n(t) \leq 1$, one can check that indeed

$$-\frac{1}{1 - F(t)} = -\frac{F(t)}{F(t)(1 - F(t)} \leq \epsilon_n \leq \frac{1 - F(t)}{F(t)(1 - F(t))} = \frac{1}{F(t)}, \tag{8}$$

so the range restraint on ϵ for the family (2) always holds for the maximum likelihood estimator, meaning that $f_{\epsilon_n}(\cdot)$ is a proper density. Now, the resulting CDF at t for this density is then,

$$\begin{aligned} F_{\epsilon_n}(t) &= \int_{-\infty}^{t} f_{\epsilon_n}(o)do \\ &= \int_{-\infty}^{t} (1 + \epsilon_n[I(o \leq t) - F(t)])f(o)do \\ &= \int_{-\infty}^{t} f(o)do + \epsilon_n \int_{-\infty}^{t} I(o \leq t)f(o)do - \epsilon_1 F(t) \int_{-\infty}^{t} f(o)do \end{aligned}$$

$$
\begin{aligned}
&= F(t) + \epsilon_n F(t) - \epsilon_n F(t)^2 = F(t) + \epsilon_1 F(t)(1 - F(t)) \\
&= F(t) + \frac{F_n(t) - F(t)}{F(t)(1 - F(t))} F(t)(1 - F(t)) \text{ from (7)} \\
&= F(t) + F_n(t) - F(t) = F_n(t).
\end{aligned}
$$

Therefore, for any initial density $f(\cdot)$ and any time point t, the targeted likelihood maximum likelihood estimator of the CDF reduces to the empirical distribution estimator in a single step. This result immediately generalizes to $\Psi(p) = \int_A p(o)d\mu(o)$ for any measurable set A.

3 Solving the efficient estimating equation.

We have the following trivial, but useful result. It states that if the MLE's $\epsilon(P_n \mid p_n^k)$ at step k of the targeted MLE algorithm converge to zero for $k \to \infty$ (as one expects to hold if the log likelihood of the data is uniformly bounded in the model $\mathcal{M}$), then the algorithm converges to a solution of the efficient influence curve equation $P_n D(p) = 0$ in the sense that $P_n D(p_n^k) \to 0$.

Result 1 *Let P_n be given. Assume that*

$$
\lim_{\epsilon \to 0} \limsup_{k \to \infty} \mid P_n \frac{\frac{d}{d\epsilon} p_n^k(\epsilon)}{p_n^k(\epsilon)} - P_n \frac{p_n^{k\prime}(0)}{p_n^k(0)} \mid \to 0, \tag{9}
$$

that for each k there exist a constant matrix A_k so that $A_k \frac{p_n^{k\prime}}{p_n^k} = D(p_n^k)$ with $\limsup_{k \to \infty} \parallel A_k \parallel < \infty$, where $\parallel A \parallel$ denotes a matrix norm.

If $\epsilon(P_n \mid p_n^k)$ solves $P_n \frac{\frac{d}{d\epsilon} p_n^k(\epsilon)}{p_n^k(\epsilon)} = 0$ for all k, and $\epsilon(P_n \mid p_n^k) \to 0$ for $k \to \infty$, then we have

$$
P_n D(p_n^k) \to 0 \text{ for } k \to \infty.
$$

The condition (9) holds if the score of the one-dimensional submodel $p(\epsilon)$ at ϵ converges to the score at $\epsilon = 0$ for $\epsilon \to 0$ uniformly in a set containing the k-step targeted MLE's p_n^k, $k = 1, 2, \ldots$, and that for each $p \in \mathcal{M}$, the linear span of the components $\frac{p'(0)}{p(0)}$ includes the components of $D(p)$. Since the likelihood increases at each step one might indeed expect that typically the targeted MLE algorithm will converge and thereby that $\epsilon(P_n \mid p_n^k) \to 0$. That is, Result 1 essentially states that, if the targeted MLE algorithm converges, then the algorithm will converge to a solution of the efficient influence curve equation in the sense that by choosing k large enough $P_n D(p_n^k) \approx 0$ with

arbitrary small deviation from 0.

Proof. Let $\epsilon_k = \epsilon(P_n \mid p_n^k)$, $k = 0, \ldots$. If $\epsilon_k \to 0$ for $k \to \infty$, then

$$P_n \frac{\frac{d}{d\epsilon_k} p_n^k(\epsilon_k)}{p_n^k(\epsilon_k)} - P_n \frac{p_n^{k\prime}(0)}{p_n^k(0)} \to 0$$

for $k \to \infty$. Let A_k be such that $A_k \frac{p_n^{k\prime}(0)}{p_n^k(0)} = D(p_n^k)$. By assumption, the matrix has a norm bounded uniformly in k. Thus, we also have

$$P_n A_k \frac{\frac{d}{d\epsilon_k} p_n^k(\epsilon_k)}{p_n^k(\epsilon_k)} - P_n D(p_n^k) \to 0$$

for $k \to \infty$. However, $P_n \frac{d}{d\epsilon_k} p_n^k(\epsilon_k)/p_n^k(\epsilon_k) = 0$ (and thus A_k applied to this equals 0 as well), which shows that $P_n D(p_n^k) \to 0$. □

4 Efficiency of targeted likelihood estimation.

In this section we provide templates for proving consistency, asymptotic linearity and efficiency of the targeted maximum likelihood estimator of a path-wise differentiable parameter. Since convexity of the model and linearity of the parameter allows a particular strong result, we separate this situation from the general case.

4.1 Linear parameters in convex models.

Let p_n^∞ denote the limit of our algorithm if it exists as a density with respect to μ in $\mathcal{M}$, and otherwise it represents a $p_n^k \in \mathcal{M}$ for a large enough k. If the condition of the above Result 1 holds, then $p_n^\infty \in \mathcal{M}$, and for all practical purposes, we have $P_n D(p_n^\infty) = 0$. If this is true, then this result can be used to establish efficiency of the substitution estimator $\Psi(p_n^\infty)$ as an estimator of ψ_0 under the assumption that the parameter $\Psi : \mathcal{M} \to \mathbb{R}^d$ is linear and $\mathcal{M}$ is convex, under weak regularity conditions. Specifically, by the identity for convex models and linear parameters in van der Laan (1998) we have $\Psi(p) - \Psi(p_0) = -P_0 D(p)$ for any $p, p_0 \in \mathcal{M}$ for which $p_0/p < \infty$. Thus, if $p_n^\infty \in \mathcal{M}$ and it is bounded away from 0 on the support of p_0, then combining $P_n D(p_n^\infty) = 0$ with the latter identity gives us

$$\Psi(p_n^\infty) - \Psi(p_0) = (P_n - P_0) D(p_n^\infty). \tag{10}$$

Even if p_n^∞ does not satisfy $p_0/p_n^\infty < \infty$, then the identity $\Psi(p_n^\infty) - \Psi(p_0) = -P_0 D(p_n^\infty)$ can still be established under a continuity condition on $p \to P_0 D(p)$

(see van der Laan (1998)), so that (10) can even be established for density estimators not satisfying this support condition.

Applying empirical process theory (van der Vaart and Wellner (1996a)) now proves that $\Psi(p_n^\infty)$ is root-n consistent if $D(p_n^\infty)$ falls in a P_0 Donsker class with probability tending to 1. If one can now also establish that $P_0(D(p_n^\infty)-D(p_1))^2$ converges to zero in probability for a certain $p_1 \in \mathcal{M}$, then it follows that $\Psi(p_n^\infty)$ is asymptotically linear with influence curve $D_0(p_1) \equiv D(p_1)-P_0D(p_1)$:

$$\Psi(p_n^\infty) - \Psi(p_0) = (P_n - P_0)D_0(p_1) + o_P(1/\sqrt{n}),$$

where we note that p_1 can be an arbitrary limit (i.e., $p_1 \neq p_0$ is allowed). In particular, if the limit p_1 is such that $D(p_1) = D(p_0)$, then $\Psi(p_n^\infty)$ is asymptotically linear with influence curve $D(p_0)$. Thus, if $D(p_0)$ is the efficient influence curve, then $\Psi(p_n^\infty)$ is asymptotically efficient.

Theorem 1 *Suppose the conclusion of Result 1 holds, and $K = K(n)$ is chosen large enough so that the targeted MLE $p_n = p_n^K$ satisfies $P_nD(p_n) = R(n, K(n)) = o_P(1/\sqrt{n})$ (where $\lim_{K\to\infty} R(n, K) = 0$). Assume that $p_n \in \mathcal{M}$, $p_0/p_n < \infty$ uniformly over a support of p_0, $\mathcal{M}$ is convex, and $\Psi : \mathcal{M} \to \mathbb{R}^d$ is linear. Then*

$$\Psi(p_n) - \Psi(p_0) = (P_n - P_0)D(p_n) + R(n, K(n)).$$

If $D(p_n)$ falls in a P_0 Donsker class with probability tending to 1, then

$$\Psi(p_n) - \psi_0 = O_P(1/\sqrt{n}).$$

If it is also shown that $P_0(D(p_n) - D(p_1))^2 \to 0$ in probability for $n \to \infty$ for some $p_1 \in \mathcal{M}$, then it follows that $\Psi(p_n)$ is asymptotically linear with influence curve $D(p_1) - P_0D(p_1)$:

$$\Psi(p_n) - \Psi(p_0) = (P_n - P_0)D(p_1) + o_P(1/\sqrt{n}).$$

In particular, if $D(p_1) = D(p_0)$, and $D(p_0)$ is the efficient influence curve of Ψ at p_0, then $\Psi(p_n)$ is asymptotically efficient.

This shows that the targeted MLE of a linear parameter in a convex model is typically consistent and asymptotically linear for arbitrary starting density p_n^0, and if the targeted MLE p_n^∞ is consistent in the sense that $P_0(D(p_n^\infty) - D(p_0))^2 \to 0$ with probability tending to 1 for n converging to infinity (e.g., the initial starting density p_n^0 would already yield a consistent estimator $D(p_0^n)$ of $D(p_0)$), then the targeted MLE will also be efficient. We will now provide

two examples illustrating this theorem. The first example represents a case in which the targeted MLE is efficient for arbitrary starting density p_n^0. The second example represents the case that the targeted MLE is consistent and asymptotically linear for arbitrary starting density p_n^0, and is efficient if the starting density consistently estimates $D(p_0)$.

Example 1 ((Efficiency of a smooth cumulative distribution function) In this example we have $D(p)(O) = I(O \leq t) - \int_0^t p(o)d\mu(o)$. A targeted MLE p_n solving $P_n D(p_n) = 0$ satisfies that $\Psi(p_n) = P_n I(\cdot \leq t)$ equals the empirical cumulative distribution function at t and is therefore asymptotically efficient, *for arbitrary starting density* p^0. Thus in this example the initial density does not need to be consistent in order to make the targeted MLE asymptotically efficient. Suppose that p_{nh}^0 is indexed by a bandwidth or model choice h, and let p_{nh}^* be the targeted MLE density estimator using as starting density p_{nh}^0. Each of the targeted MLE's p_{nh}^* results in the same estimator of the cumulative distribution function $\Psi(p_0)$ at time t. If one uses likelihood cross-validation to select h, then one selects among all of these targeted MLE's the one which is supposedly closest to the true density p_0 with respect to Kullback-Leibler divergence, which now provides a valid and reasonable criteria since all the candidates density estimators already map into efficient (and algebraically equivalent) estimators of ψ_0.

Example 2 ((Local efficiency of targeted MLE based on censored data) We consider a particular example of a censored data structure to illustrate that Theorem 1 yields local efficiency of the targeted MLE based on CAR censored data structures based on any starting density p_n^0, under very weak conditions.

Suppose that the full data structure $X = (W, Y(a) : a \in \{0,1\})$ on the experimental unit consists of a set of baseline covariates W, and treatment specific outcomes $Y(a)$, indexed by treatment values $a \in \{0,1\}$. Suppose that the observed data structure $O = (W, A, Y = Y(A)) \sim p_0$, and it is assumed that the conditional probability distribution $g_0(\cdot \mid X)$ of A, given X, satisfies $g_0(A \mid X) = g_0(A \mid W)$: that is, A is independent of X, given W. Suppose that this conditional probability distribution of $g_0(A \mid W)$ of A, given W, is known, and satisfies $0 < g_0(1 \mid W) < 1$, as it would be in a randomized trial aiming to establish the causal effect of A on Y. Let $\mathcal{M}$ be the class of all densities of O with respect to an appropriate dominating measure. We have

$$\mathcal{M} = \{p(O) = Q_{XA}(W, Y)g_0(A \mid X) : Q_{X0}, Q_{X1}\},$$

where the full data sub-distributions $Q_{Xa}(w, y) = P_{W,Y(a)}(w, y)$ are joint densities of $(W, Y(a))$, $a \in \{0,1\}$, and are unspecified. As a consequence, $\mathcal{M}$ is

a convex model. Let $\Psi : \mathcal{M} \to \mathbb{R}$ be defined as $\Psi(p) = E_p(Y(1) - Y(0)) = E_p(E_p(Y \mid A = 1, W) - E_p(Y \mid A = 0, W))$, which is often called the marginal causal effect of treatment A on the outcome Y. In this case, $\Psi(p)$ is pathwise differentiable at p with efficient influence curve $S(p)$ defined by

$$S(p) = \frac{(Y - Q(p)(A, W))(A - (1 - A))}{g(p)(A \mid W)} + Q(p)(1, W) - Q(p)(0, W) - \Psi(p),$$

where $g(p)(\cdot \mid W) = Pr_p(A = \cdot \mid W) = g_0(\cdot \mid W)$, and $Q(p)(A, W) = E_p(Y \mid A, W)$. Note that $\Psi(p)$ depends on p through $Q(p)$ and its marginal distribution p_W of W. Due to the factorization of the density of O in a Q_X-factor and g_0 factor, this is also the efficient influence curve if g_0 is unknown or modelled. The class of all gradients at $p \in \mathcal{M}$ is given by:

$$\left\{ \frac{(Y - Q(A, W))(I(A = 1) - I(A = 0))}{g_0(A \mid W)} + Q(1, W) - Q(0, W) - \Psi(p) : Q \right\},$$

where Q can be an arbitrary function of A, W.

So we could define

$$D_Q(p)(O) \equiv \frac{(Y - Q(A, W))(A - (1 - A))}{g_0(A \mid W)} + Q(1, W) - Q(0, W) - \Psi(p),$$

and $D(p) = D_{Q(p)}(p)$ represents the efficient influence curve. We are now ready to define the targeted MLE of p_0 with respect to the parameter ψ_0.

Let p_n^0 be an initial density estimator of p_0. For example, p_n^0 could correspond with the empirical distribution of W, and a normal distribution for the conditional density of Y, given A, W, with mean $Q_n^0(A, W)$ and variance $\sigma_n^2(A, W)$, where Q_n^0 is an estimate of $Q(p_0)(A, W) = E_0(Y \mid A, W)$. Let p_n^* be a targeted MLE, as we explicitly define in the later Section 6 in detail, solving $P_n D(p_n^*) = 0$. In Section 6, we show for a particular hardest submodel $p_n^k(\epsilon)$ consisting of normal densities of Y, conditional on A, W, with ϵ corresponding with a fluctuation of current regression $Q_n^k(A, W)$, that the targeted MLE is achieved in the first step (i.e., $p_n^* = p_n^1$), and indeed solves the score equation $P_n D(p_n^1) = 0$. Let's consider this particular targeted MLE for illustration, but the following arguments apply to any targeted MLE solving $P_n D(p_n^*) = 0$.

Application of the theorem teaches us that

$$\Psi(p_n^*) - \psi_0 = (P_n - P_0) D_{Q(p_n^*)}.$$

Since g_0 is bounded away from zero, if Q_n^1 is a nice smooth function (e.g., with a uniformly bounded uniform sectional variation norm, van der Laan

(1995)), it follows that $D_{Q(p_n^*)}$ falls in a P_0-Donsker class, and thus that $\Psi(p_n^*) - \psi_0 = O_P(1/\sqrt{n})$. If the initial regression estimator $Q_n^0 = Q(p_n^0)$ converges to a possibly misspecified $Q_1 = Q(p_1)$, then it follows that $\Psi(p_n^*)$ is asymptotically linear with influence curve $D_{Q(p_1)}(O)$, where p_1 is the possibly misspecified limit of p_n^1. Finally, if Q_n^0 is actually consistent for $Q(p_0)$, then the targeted MLE of ψ_0 is asymptotically efficient. We can use likelihood based cross-validation to select among targeted MLE's indexed by different candidate initial estimators Q_n^0, thereby improving the efficiency relative to a targeted MLE with a fixed initial Q_n^0. Thus this example teaches us that the targeted MLE $\Psi(p_n^*)$ of ψ_0, which typically equals the first step targeted MLE, is consistent and asymptotically linear for arbitrary initial regression estimator Q_n^0, and it is efficient if Q_n^0 happens to be consistent, where the latter can potentially be achieved by using a machine learning type algorithm and selecting the fine tuning parameters with likelihood based cross-validation. These results still carry through if g_0 is unknown but is known to belong to a parametric model.

4.2 Local efficiency for general smooth parameters.

The remarkable robustness with respect to the starting density p_n^0 as observed in the previous subsection is a consequence of the convexity of the model and linearity of the parameter Ψ. In general, such results cannot be expected to hold. In this subsection we present a more general approach for establishing the wished asymptotic linearity and efficiency of the targeted MLE of any pathwise differentiable parameter.

Let $p_n^\infty \in \mathcal{M}$ denote the limit of the targeted MLE algorithm if it exists and otherwise it represents a p_n^k for a large k. If the targeted MLE solves the efficient influence curve equation, then for all practical purposes, we have $P_n D(p_n^\infty) = 0$. Let $R(p, p_0)$ be defined by

$$\Psi(p) - \Psi(p_0) = -P_0 D(p) + R(p, p_0)$$

for any $p \in \mathcal{M}$. We note that by pathwise differentiability of Ψ at p, $R(p, p_0)$ represents a second order term in the difference $p - p_0$. Combining $P_n D(p_n^\infty) = 0$ with the latter identity gives us

$$\Psi(p_n^\infty) - \Psi(p_0) = (P_n - P_0) D(p_n^\infty) + R(p_n^\infty, p_0).$$

Applying empirical process theory now proves that $\Psi(p_n^\infty)$ is root-n consistent if $D(p_n^\infty)$ falls in a P_0 Donsker class with probability tending to 1, and $R(p_n^\infty, p_0) = o_P(1/\sqrt{n})$. If one can now also establish that $P_0(D(p_n^\infty) - D(p_1))^2$

converges to zero in probability for a possibly misspecified $p_1 \in \mathcal{M}$, then it follows that $\Psi(p_n^\infty)$ is asymptotically linear with influence curve $D(p_1) - P_0 D(p_1)$:

$$\Psi(p_n^\infty) - \Psi(p_0) = (P_n - P_0)D(p_1) + o_P(1/\sqrt{n}).$$

In particular, if $D(p_1) = D(p_0)$, then the targeted MLE is asymptotically efficient. Note that the asymptotic linearity requires that $R(p_n^\infty, p_0) = o_P(1/\sqrt{n})$, while the convexity of the model and linearity of the parameter as assumed in the previous subsection allowed us to avoid such a condition: i.e. in that case we had $R(p, p_0) = 0$ for arbitrary $p \in \mathcal{M}$ with $p_0/p < \infty$.

5 Fusion of MLE and estimating equations

In this section we show that the targeted MLE can be viewed as a solution of an optimal estimating equation for the parameter of interest, if one estimates the nuisance parameters with the targeted MLE itself. This comparison can only be made by making the assumption that the efficient influence curve can be viewed as an estimating function of the parameter of interest, which is needed for the estimating function methodology (van der Laan and Robins (2003)), but not for targeted MLE.

As previously argued, a sieve-based maximum likelihood estimator of a pathwise differentiable parameter is based on choices such as the sieve and the criteria for trading off variance and bias, which is completely unrelated to the actual parameter Ψ. As a consequence, such likelihood based estimators suffer, in principle, from serious bias for the parameter of interest ψ_0. Let p_n^0 be such a likelihood based estimator of p_0 and $\Psi(p_n^0)$ be the corresponding substitution estimator of ψ_0.

On the other hand, estimating function methodology (van der Laan and Robins (2003)) constructs estimating functions $D_h(\psi, \upsilon)(O)$ for the parameter of interest ψ indexed by a choice h, based on a representation of the orthogonal complement of the nuisance tangent space $p \to T_{nuis}^\perp(p)$ (i.e., $D_h(\Psi(p), \Upsilon(p)) \in T_{nuis}^\perp(p)$ for all h), which typically also depend on an unknown nuisance parameter Υ satisfying $E_p D_h(\Psi(p), \Upsilon(p)) = 0$ for all $p \in \mathcal{M}$. The current recommendation in estimating function methodology (see e.g., van der Laan and Robins (2003)) proposes to use an external estimator υ_n of nuisance parameters and estimate ψ_0 with the solution of $0 = P_n D_{h_n}(\psi, \upsilon_n) = 0$ in ψ. For example, one could use the maximum likelihood estimator p_n^0 and estimate ψ_0 with the solution ψ_{n0} of $0 = P_n D_{h(p_n^0)}(\psi, \Upsilon(p_n^0))$. This estimator ψ_{n0} is not necessarily, and in fact, will typically not be equal to $\Psi(p_n^0)$. Thus, even if the nuisance parameters are based on a maximum likelihood estimator p_n^0, the

resulting estimating function based estimators of ψ_0 are intrinsically different from (and less biased than) the likelihood based estimator $\Psi(p_n^0)$.

However, let p_n be the targeted maximum likelihood estimator based on hardest submodels at p with efficient influence curve $D(p) = D_{h(p)}(\Psi(p), \Upsilon(p))$ and starting with the initial density estimator p_n^0, so that p_n solves $P_n D(p_n) = D_{h(p_n)}(\Psi(p_n), \Upsilon(p_n)) = 0$. Again, we consider the (now targeted) maximum likelihood estimator $\Psi(p_n)$ versus the estimating function based estimator described in the previous paragraph. The estimating function based estimator ψ_n of ψ_0 is defined as the solution of the estimating equation $0 = P_n D_{h(p_n)}(\psi, \Upsilon(p_n))$, which differs from above by now using the targeted MLE p_n (based on p_n^0) to estimate the index and nuisance parameters (instead of likelihood based p_n^0). Because $P_n D_{h(p_n)}(\Psi(p_n), \Upsilon(p_n)) = 0$, it follows that the estimating function based estimator ψ_n now equals $\Psi(p_n)$, assuming that this solution is unique. That is, if one estimates the nuisance parameters and index in the estimating function methodology with a targeted maximum likelihood estimator p_n, then the (or, at least, one of the) estimating function based estimator ψ_n and the targeted maximum likelihood estimator $\Psi(p_n)$ are identical.

Note that the targeted MLE is more general than the estimating function based methodology since it does not require the representation of an estimating function as a function of the parameter of interest and a variation independent nuisance parameter, thereby making it more widely applicable. Another advantage of targeted MLE relative to estimating function based estimation that it is invariant to monotone transformations of the parameter of interest.

5.1 CAR-censored data models

This targeted MLE approach has a particular nice application in estimation of pathwise differentiable parameters based on censored data under the coarsening at random assumption (Heitjan and Rubin (1991), Jacobsen and Keiding (1995), Gill et al. (1997), van der Laan and Robins (2003)). That is, let $O = \Phi(C, X) \sim p_0$ for some known many to one mapping Φ, $X \sim F_{X0}$ is the full data structure one wishes to observe on a randomly sampled experimental unit, and assume that the conditional distribution of the censoring variable C, given X, i.e., the censoring mechanism, satisfies coarsening at random (CAR). In this case it is known that the density of O factorizes as: $p_0(0) = g(p_0)(O \mid X)Q(p_0)(O)$, where $g(p_0)(O \mid X)$ (which is only a function of O by CAR) is the conditional density of O, given X, which thus only depends on the conditional distribution of C, given X. The $Q(p_0)$ factor only depends on the distribution F_{X0} of the full data structure X (van der Laan and Robins (2003)). Thus given a model $\mathcal{M}$ for O obtained by modelling

F_{X0} and or the censoring mechanism $g_0(O \mid X)$, each $p \in \mathcal{M}$ is identified by $(g(p), Q(p))$. Let $\Psi(p) = \Psi(Q(p))$ be a pathwise differentiable parameter of the $Q(p)$-part of the density p of O: i.e., it represents an identifiable parameter of F_X. In this case, it is known that the efficient influence curve $D(p) = D(g(p), Q(p))$ at $p \in \mathcal{M}$ is orthogonal to the tangent space $T_{CAR}(p)$ of the censoring mechanism g at p only assuming CAR (i.e., the Hilbert space in $L_0^2(P)$ spanned by all scores of parametric submodels through $g(p)$ at p), where $T_{CAR}(p) = \{h(O) : E_p(h(O) \mid X) = 0\}$ consists of all functions of O with conditional mean, given X, equal to zero. As a consequence, given an initial estimator Q^0 of $Q(p_0)$ and g^0 of $g(p_0)$, a hardest parametric model for ψ_0 can be chosen to be of the form $p^0(\epsilon) \approx (1 + \epsilon D(p^0))p^0 = g^0 Q^0(\epsilon)$, where $Q^0(\epsilon) \approx (1 + \epsilon D(Q^0, g^0))Q^0$. That is, the hardest parametric model only corresponds with changing Q^0, but it leaves g^0 untouched. The targeted MLE approach proceeds now as defined above.

5.2 Targeting the censoring mechanism.

In this subsection we propose a targeted maximum likelihood methodology for estimation of ψ_0 which involves updating of estimators of both g_0 and Q_0. As shown in van der Laan and Robins (2003) (Theorem 1.3), we have that any gradient $D(p)$ can be decomposed as $D(p) = D_{IPCW}(p) - D_{CAR}(p)$ with D_{IPCW} being a so called Inverse Probability of Censoring Weighted (IPCW) function, and $D_{CAR}(p) = \Pi(D_{IPCW}(p) \mid T_{CAR}(p))$ is the projection of the IPCW function $D_{IPCW}(p)$ onto $T_{CAR}(p)$ in the Hilbert space $L_0^2(p)$. In order to relate these functions to estimating functions for ψ_0 (as in van der Laan and Robins (2003)) we will also sometimes use $D_{IPCW}(p) = D_{IPCW}(g(p), \Psi(p))$ and $D(p) = D(g(p), Q(p), \Psi(p))$ in the case that these functions can be represented as an estimating function in ψ indexed by nuisance parameters being functions of $g(p)$ and $Q(p)$: we note that the IPCW estimating function typically only depends on p through $g(p)$ and $\Psi(p)$. Given an initial estimator $p_n^0 = (g_n^0, Q_n^0)$, in the censored data literature one defines the IPCW-estimator and DR-IPCW estimator as the solutions of the estimating equations $P_n D_{IPCW}(g_n^0, \psi) = 0$ and $P_n D(g_n^0, Q_n^0, \psi) = 0$, respectively, and $\Psi(Q_n^0)$ is called the likelihood based estimator (making the assumption that Q_n^0 is likelihood based).

We will now describe the targeted MLE algorithm also involving the updating of g_n^0. At step k it now involves also a parametric submodel $g(p_n^k)(\epsilon_2)$ through $g(p_n^k)$ with score $D_{CAR}(g_n^k, Q_n^k)$ at $\epsilon_2 = 0$. It can be shown that $D_{CAR}(g(p), Q(p))$ corresponds with the efficient influence curve of the parameter $\Phi(g) = E_p D_{IPCW}(g, Q(p))$ at $g = g(p)$, so that this parametric submodel

makes the estimator of g_0 targeted for estimation of the mean of the $IPCW$-component of the efficient influence curve. In particular, it is also the parametric submodel which makes the IPCW estimator $\psi_{n,IPCW}$, defined as the solution of the IPCW estimating equation $0 = P_n D_{IPCW}(g_n, \psi)$, efficient if the submodel is correctly specified, under regularity conditions. As above, let $Q_n^k(\epsilon_1)$ be a parametric submodel through Q_n^k with score $D(g_n^k, Q_n^k)$ at $\epsilon_1 = 0$.

Targeted MLE algorithm:

- Set $k = 0$.
- Let $p_n^k = (g_n^k, Q_n^k)$.
- Let $\epsilon_{1nk} = \arg\max_{\epsilon_1} P_n \log Q_n^k(\epsilon_1)$, and $\epsilon_{2nk} = \arg\max_{\epsilon_2} P_n \log g_n^k(\epsilon_2)$.
- Set $g_n^{k+1} = g_n^k(\epsilon_{2n})$ and $Q_n^{k+1} = Q_n^k(\epsilon_{1n})$. Set $p_n^{k+1} = (g_n^{k+1}, Q_n^{k+1})$.
- Set $k = k + 1$, and iterate this process utill convergence.

If ϵ_{1nk} and ϵ_{2nk} converge to zero for $k \to \infty$ (which can be expected because both factors g and Q of the likelihood are increasing at each step), then the targeted MLE algorithm will converge to a simultaneous solution of

$$\lim_k P_n D_{CAR}(g^k, Q^k) = 0 \text{ and } \lim_k P_n D(g^k, Q^k) = 0.$$

Equivalence of IPCW, DR-IPCW, and targeted MLE: As a consequence of the decomposition $D(p) = D_{IPCW}(p) - D_{CAR}(p)$, this implies also $\lim_k D_{IPCW}(g^k, \Psi(Q^k)) = 0$. Note that the double robust IPCW estimator defined as the solution in ψ of $P_n D(g_n^k, Q_n^k, \psi) = 0$, the targeted maximum likelihood estimator $\Psi(Q_n^k)$, and the IPCW estimator defined as the solution of $P_n D(g_n^k, \psi) = 0$, all based on these targeted MLE's g_n^k, Q_n^k are identical up to an arbitrarily small error decreasing in k (assuming uniqueness of the DR-IPCW and IPCW solution).

6 Examples of targeted maximum likelihood.

In this section we provide some important examples of the targeted MLE to illustrate its remarkable simplicity and good properties. For additional examples we refer to our accompanying technical report.

6.1 Estimation of a mean in a nonparametric model.

Consider an initial data generating density p_n^0 (with respect to a dominating measure μ) of a possibly multivariate random variable O, a given function $w(\cdot)$, and define the parameter of interest as

$$\Psi(p) = E_p[w(O)] = \int w(o)p(o)d\mu(o).$$

For the exponential family

$$\left\{ p_n^0(\epsilon)(x) = \frac{\exp(\epsilon(w(x) - \psi_n^0))p_n^0(x)}{\int \exp(\epsilon(w(x) - \psi_n^0))p_n^0(x)d\mu(x)} : \epsilon \right\},$$

consider attempting to estimate ϵ with maximum likelihood based on an i.i.d. sample $\{O_i\}_{i=1}^n$. Here $\psi_n^0 = \Psi(p_n^0)$. The log likelihood is then,

$$l(\epsilon) = \sum_{i=1}^{n} [\log(p_n^0(O_i)) + \epsilon(w(O_i) - \psi_n^0) - \log\left(\int \exp(\epsilon(w(x) - \psi_n^0))p_n^0(x)d\mu(x)\right)].$$

In our accompanying technical report we show that (for each initial p_n^0) the one-step targeted maximum likelihood estimator $\Psi(p_n^1) = \Psi(p_n^0(\epsilon_n)$ of the mean of $w(O)$ equals the sample mean $\bar{W}_n = \frac{1}{n}\sum_{i=1}^n w(O_i)$. For the detailed proof we refer to our technical report.

6.2 Estimation of a marginal causal effect.

Double robust locally efficient estimation of the causal effect of a point treatment assuming a marginal structural model has been provided in Robins (2000), Robins and Rotnitzky (2001), and Robins et al. (2000): see also van der Laan and Robins (2003).

Let $O = (W, A, Y)$, W be a vector of baseline covariates, A be a binary treatment variable, and Y an outcome of interest. Let $\mathcal{M}$ be the class of all densities of O with respect to an appropriate dominating measure: so $\mathcal{M}$ is nonparametric up to possible smoothness conditions. Let $\Psi : \mathcal{M} \to \mathbb{R}$ be defined as $\Psi(p) = E_p(E_p(Y \mid A = 1, W) - E_p(Y \mid A = 0, W))$, where it is assumed $0 < P(A = 1 \mid W) < 1$ with probability one so that this parameter is well defined. This parameter corresponds with the marginal causal effect of A on Y if one assumes the usual consistency assumption, temporal ordering assumption, and randomization assumption required for causal inference. In order to acknowledge that this parameter is of interest in general, van der Laan (2006) refers to this parameter as the variable importance of variable

A. This parameter $\Psi(p)$ is pathwise differentiable at p with efficient influence curve $S(p)$ defined by

$$\begin{aligned} S(p) &= \frac{(Y - Q(p)(A,W))(I(A=1) - I(A=0))}{g(p)(A \mid W)} \\ &\quad + Q(p)(1,W) - Q(p)(0,W) - \Psi(p), \end{aligned}$$

where $g(p)(\cdot \mid W) = Pr_p(A = \cdot \mid W)$, and $Q(p)(A,W) = E_p(Y \mid A,W)$ (see e.g., Robins (2000), van der Laan (2006)). Note that $\Psi(p)$ depends on p through $Q(p)$ and its marginal distribution p_W of W. Because the model is locally saturated, it is also the *only* influence curve/gradient (Gill et al. (1997)). So we set $D(p) = S(p)$.

We can decompose this efficient score $D(p)$ into three subcomponents as follows:

$$\begin{aligned} D(p) &= D(p) - E_p(D(p) \mid A,W) + E_p(D(p) \mid A,W) - E_p(D(p) \mid W) \\ &\quad + E_p(D(p) \mid W) - E_p D(p), \end{aligned}$$

which corresponds with scores for $p(Y \mid A,W)$, $p(A|W)$ and $p(W)$, respectively. We have

$$\begin{aligned} D_1(p)(O) &\equiv D(p) - E_p(D(p) \mid A,W) \\ &= (Y - Q(p)(A,W))\frac{A - (1-A)}{g(p)(A \mid W)} \\ E_p(D(p) \mid A,W) - E_p(D(p) \mid W) &= 0 \\ D_2(p) &\equiv E_p(D(p) \mid W) - E_p(D(p)) \\ &= Q(p)(1,W) - Q(p)(0,W) - \Psi(p). \end{aligned}$$

Consider an initial density estimator p_n^0 of the density p_0 of (W,A,Y) with marginal distribution of W being the empirical probability distribution of $W_1,\ldots,W_n$. We have that $D(p_n^0) = D_1(p_n^0) + D_2(p_n^0)$ and thus that a one-dimensional $p_n^0(\epsilon)$ with score $D(p_n^0)$ at $\epsilon = 0$ corresponds with a zero score for $g(p_n^0)$. In addition, we have that $P_n D_2(p_n^0) = 0$ (i.e., the empirical distribution of W is a nonparametric maximum likelihood estimator) so that $p_n^0(\epsilon)$ can be selected to only vary $p_n^0(Y \mid A,W)$ with a score $D_1(p_n)$ at $\epsilon = 0$.

We now propose an easily implemented targeted maximum likelihood estimator of the marginal causal effect by using a normal regression model as hardest submodel. Specifically, consider an initial density estimator p_n^0 with marginal distribution of W equal to the empirical probability distribution of $W_1,\ldots,W_n$, and let the conditional probability density $p_n^0(Y \mid A,W) =$

$\frac{1}{\sigma(Q_n^0)(A,W)} f_0(\{Y - Q_n^0(A,W)\}/\sigma(Q_n^0)(A,W))$ be a normal density with mean $Q_n^0(A,W)$ and variance $\sigma(Q_n^0)^2(A,W)$. Here f_0 denotes the $N(0,1)$ density. In addition, $g(p_0^n)(A \mid W)$ is a particular fit of the conditional density of A, given W. We now consider as possible submodels $p_n^0(\epsilon)$

$$p_n^0(\epsilon)(Y \mid A,W) = \frac{1}{\sigma(Q_n^0(A,W)} f_0\left(\frac{Y - Q_n^0(A,W) - \epsilon h(p_n^0)(A,W)}{\sigma(Q_n^0)(A,W)}\right),$$

where the function h will be specified so that the score of p_n^0 at $\epsilon = 0$ equals the efficient influence curve at p_n^0. The maximum likelihood estimator of ϵ is simply given by the weighted least squares estimator for a univariate linear regression model:

$$\epsilon_n = \arg\min_{\epsilon} \sum_{i=1}^{n} (Y_i - Q_n^0(A_i,W_i) - \epsilon h(p_n^0)(A_i,W_i))^2 \frac{1}{\sigma(Q_n^0)^2(A_i,W_i)}.$$

The score of $p_n^0(\epsilon)(Y \mid A,W)$ at a value ϵ is given by:

$$S(\epsilon) = -\frac{Y - Q_n^0(A,W) - \epsilon h(p_n^0)(A,W)}{\sigma(Q_n^0)^2(A,W)} h(p_n^0)(A,W),$$

and ϵ_n solves indeed $P_n S(\epsilon_n) = 0$. If we set

$$h(p_n^0)(A,W) \equiv \left(\frac{I(A=1)}{g_n^0(1 \mid W)} - \frac{I(A=0)}{g_n^0(0 \mid W)}\right) \sigma(Q_n^0)^2(A,W),$$

then the score $S(0) = D_1(p_n^0) = (Y - Q_n^0(A,W))(I(A=1)/g_n^0(1 \mid W) - I(A=0)/g_n^0(0 \mid W))$ of $p_n^0(\epsilon)(Y \mid A,W)$ at $\epsilon = 0$ corresponds with the efficient influence curve at p_n^0. As in our previous subsection, since $p_n^0(W)$ equals the empirical distribution of W the MLE of $\epsilon_1 \rightarrow P_n \log p^0(\epsilon_1)(W)$ equals $\epsilon = 0$, and $g_n^0(A \mid W)$ will not be varied by $p_n^0(\epsilon)$: that is, the marginal distribution of W and the treatment mechanism $g^0(A \mid W)$ will not be updated in the algorithm for calculating the targeted maximum likelihood estimator.

Let $p_n^1 = p_n^0(\epsilon_n)$ whose conditional distribution of Y, given A,W, is a normal density with mean $Q_n^1(A,W)$ and variance $\sigma^2(Q_n^1)(A,W)$, where

$$Q_n^1(A,W) = Q(p_n^1)(A,W) = Q_n^0(A,W) + \epsilon_n h(p_n^0)(A,W).$$

The corresponding estimate of ψ_0 is given by

$$\Psi(p_n^1) = \frac{1}{n} \sum_{i=1}^{n} Q_n^1(1,W_i) - Q_n^1(0,W_i).$$

It is straightforward to show that $P_n D(p_n^1) = 0$ in the case that $\sigma_n^0(A,W)$ is constant in the model $\{p_n^0(\epsilon) : \epsilon\}$, but is simply set at an initial estimate. Thus in this case the targeted maximum likelihood is achieved at the first step. For arbitrary fixed values of $\sigma(A,W)$, the targeted MLE is locally efficient in the sense that if $g(p_n^0)$ is consistent at some rate, then it is consistent and asymptotically linear for arbitrary Q_n^0, and it is efficient if Q_n^0 is consistent for $Q_0(A,W)$. Likewise, a consistent $Q_n^1(A,W)$ will lead to a consistent estimator of the parameter of interest ψ_0, even with an arbitrary fit of the treatment mechanism $g(A|W)$. Iterative estimation of σ provides no (asymptotic) reward, and could simply be omitted by setting (e.g.) σ at an initial estimate, so that the targeted MLE is achieved in a single step.

6.3 Targeting the treatment mechanism as well.

We will now proceed with this example, but also use for g_0 a targeted maximum likelihood estimator. Our goal is to make the IPTW estimator $\psi_{n,IPTW} = \frac{1}{n}\sum_{i=1}^n Y_i \frac{I(A_i=1)-I(A_i=0)}{g_n(A_i|W_i)}$ corresponding wiht the targeted MLE g_n an efficient estimator. Let $g(p_n^0)(A \mid W)$ be an initial estimator and represent it as a logistic function:

$$g(p_n^0)(1 \mid W) = \frac{1}{1+\exp(-m_n^0(W))}.$$

Consider as parametric submodel

$$g(p_n^0)(\epsilon_2)(1 \mid W) = \frac{1}{1+\exp(-m_n^0(W) - \epsilon_2 h(p_n^0)(W))}. \tag{11}$$

Let $\epsilon_{2n} = \arg\max P_n \log g(p_n^0)(\epsilon)$. In practice this can be done by fitting a logistic regression in the covariates $m_n^0(W)$ and $h(p_n^0)(W)$, setting the intercept equal to zero, and setting the coefficient in front of $m_n^0(W)$ equal to 1, and set ϵ_{2n} equal to fitted coefficient in front of $h(p_n^0)(W)$. It is also fine to refit the intercept and coefficient in front of $m_n^0(W)$, since choosing additional parameters still guarantees that the linear span of scores includes the score of $h(p_n^0)(W)$. We have

$$\left.\frac{d}{d\epsilon_2}\log g(p_n^0)(\epsilon_2)\right|_{\epsilon_2=0}(O) = h(p_n^0)(W)(A - g(p_n^0)(1 \mid W)).$$

Solving for h so that

$$h(W)(A - g(p_n^0)(1 \mid W)) \;=\; D_{CAR}(p_n^0)(O)$$

$$\begin{aligned} &= \frac{Q(p_n^0)(A,W)}{g_n^0(A \mid W)} \{I(A=1) - I(A=0)\} \\ &\quad -\{Q(p_n^0)(1,W) - Q(p_n^0)(0,W)\} \end{aligned}$$

yields the solution

$$h(p_n^0)(W) = \frac{Q(p_n^0)(1,W)}{g(p_n^0)(1 \mid W)} + \frac{Q(p_n^0)(0,W)}{g(p_n^0)(0 \mid W)}.$$

We are now ready to present the proposed targeted MLE which also targets the treatment mechanism fit.

The algorithm for targeted maximum likelihood estimation of a marginal causal effect, including the targeting of the treatment mechanism. Thus the algorithm for targeted maximum likelihood estimation of ψ_0 can be described as follows. Let $k = 0$, and let $g^0(A \mid W)$ and the regression fit $Q^0(A,W)$ of $E_0(Y \mid A,W)$ be given. Let

$$h_1^k = h_1(g^k, Q^k)(A,W) \equiv \left(\frac{I(A=1)}{g^k(1 \mid W)} - \frac{I(A=0)}{g^k(0 \mid W)} \right) \sigma(Q^k)^2(A,W)$$

and

$$h_2^k = h_2(g^k, Q^k)(W) = \frac{Q^k(1,W)}{g^k(1 \mid W)} + \frac{Q^k(0,W)}{g^k(0 \mid W)}.$$

Let $m^k(W) = \log(g^k(1 \mid W)/g^k(0 \mid W))$ so that $g^k(1 \mid W) = 1/(1 + \exp(-m^k(W))$. Consider the logistic regression model

$$g^k(\epsilon_2)(1 \mid W) = \frac{1}{1 + \exp(-m^k(W) - \epsilon_2 h_2^k(W))}.$$

Let $\epsilon_{2n}(k) = \arg\max_{\epsilon_2} P_n \log g^k(\epsilon_2)$ be the maximum likelihood estimator of this univariate logistic regression model, and let

$$\epsilon_{1n}(k) = \arg\min_{\epsilon_1} \sum_{i=1}^{n} (Y_i - Q^k(A_i,W_i) - \epsilon_1 h_1^k(A_i,W_i))^2 \frac{1}{\sigma(Q^k)^2(A_i,W_i)},$$

the univariate least squares estimator of ϵ_1.

Now, update g^k and Q^k as follows:

$$\begin{aligned} Q^{k+1}(A,W) &= Q^k(A,W) + \epsilon_{1n}(k) h_1^k(A,W) \\ m^{k+1}(A,W) &= m^k(W) + \epsilon_{2n}(k) h_2^k(W) \\ g^{k+1}(A \mid W) &= \frac{1}{1 + \exp(-m^{k+1}(W))} \end{aligned}$$

Set $k = k + 1$ and iterate this algorithm.

Equivalence of IPTW, DR-IPTW, and targeted maximum likelihood estimators. Recall that the efficient influence curve function is decomposed as $D(g,Q)(O) = D_{IPTW}(g,Q) - D_{CAR}(g,Q)$, where $D_{IPTW}(g,Q) = \frac{Y}{g(A|W)}(I(A=1) - I(A=0)) - \Psi(Q)$, and $D_{CAR}(g,Q) = \frac{Q(A,W)}{g(A|W)}(I(A=1) - I(A=0)) - (Q(1,W) - Q(0,W))$. For k converging to infinity the targeted MLE yields a final estimator g_n of the treatment mechanism and a regression fit $Q_n(A,W)$ so that the score equations of the two submodels in ϵ_1 and ϵ_2 are solved at $\epsilon_1 = \epsilon_2 = 0$:

$$P_n D(g_n, Q_n) = 0 \text{ and } P_n D_{CAR}(g_n, Q_n) = 0.$$

This implies also that

$$P_n D_{IPTW}(g_n, Q_n) = 0.$$

Thus, we can conclude that the three estimators

$$\begin{aligned}
\Psi_{n,IPTW} &= \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i}{g_n(A_i \mid W_i)}(I(A_i=1) - I(A_i=0)) \\
\Psi_{n,DR-IPTW} &= \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i}{g_n(A_i \mid W_i)}(I(A_i=1) - I(A_i=0)) \\
&- D_{CAR}(g_n, Q_n)(A_i, W_i) \\
\Psi_{n,MLE} &= \frac{1}{n}\sum_{i=1}^{n} Q_n(1,W_i) - Q_n(0,W_i)
\end{aligned}$$

are algebraically identical: $\Psi_{n,IPTW} = \Psi_{n,DR-IPTW} = \Psi_{n,MLE}$. That is, the targeted MLE $\Psi(Q_n)$ equals the IPTW and DR-IPTW estimator based on the targeted MLE (g_n, Q_n) as estimators of the nuisance parameters (g_0, Q_0) in the corresponding estimating equations. Preliminary results suggest that consistency of the resulting targeted likelihood algorithm depends on the consistency of either the g_0 or Q_0 component of the initial density estimator.

6.4 Simulation for marginal variable importance.

Simulated data can be used to illustrate the benefits of the targeted likelihood procedure. We simulated replicates of the data structure $O = (W, A, Y) \sim p_0$ representing baseline covariates, a binary treatment, and a response measurement on a subject, and attempted to estimate the causal effect of treatment A on response Y. We generated 1000 datasets of size $n = 200$ according to the following mechanism:

$$W \sim U(0,1)$$

$$\begin{aligned} A &\in \{0,1\} \\ g(1|W) &= P(A=1|W) = \frac{1}{1+\exp(-8W2+8W-1)} \\ \epsilon &\sim N(0,1), \quad \epsilon \perp (W,A) \\ Y &= AQ(1,W) + (1-A)Q(0,W) + \epsilon \\ Q(0,W) &= -\frac{2}{3},\ Q(1,W) = -(8W^2 - 8W + 1) \end{aligned}$$

Here O represented a censored data structure. The unavailable *counterfactual* data was given by,

$$X = (W, Y_0, Y_1) = (W, Q(0,W) + \epsilon, Q(1,W) + \epsilon).$$

It could be be verified that the coarsening at random assumption held, or that,

$$\{A \perp X|W\},$$

as well as the experimental treatment assignment assumption, implied by,

$$0 < 0.26 < g(1|W) < .74 < 1 \text{ with probability one.}$$

Together these assumptions made it possible to estimate the parameter,

$$\Psi(p_0) = E[Y_1] - E[Y_0] = 1,$$

representing the counterfactual mean difference between the treatment group ($A = 1$) and the control group ($A = 0$).

The standard estimators for this problem are the inverse probability of treatment (IPTW), maximum likelihood (G-computation), and doubly robust (efficient) estimators. These respectively depend on fitting either the censoring mechanism g or the nuisance parameter $Q(A,W) = E[Y|W]$, and are given as follows, where $h_g(A,W) = \frac{A}{g(1|W)} - \frac{1-A}{g(0|W)}$:

$$\begin{aligned} \Psi_{n,\text{IPTW}}(g) &= \frac{1}{n}\sum_{i=1}^{n} Y_i h_g(A_i, W_i) \\ \Psi_{n,\text{MLE}}(Q) &= \frac{1}{n}\sum_{i=1}^{n} [Q(1,W_i) - Q(0,W_i)] \\ \Psi_{n,\text{DR-IPTW}}(g,Q) &= \Psi_{n,\text{IPTW}} + \Psi_{n,\text{MLE}} - \frac{1}{n}\sum_{i=1}^{n} h_g(A_i,W_i)Q(A_i,W_i) \end{aligned}$$

Typically estimation is based on forming external estimates of at least one of the two nuisance parameters g or Q, and then applying one of the IPTW,

maximum likelihood, or double robust estimators. The three estimators can potentially be very different from one another, leading to difficulties when interpreting the data. Targeted likelihood resolves this problem, by estimating both nuisance parameters g and Q accurately with maximum likelihood, but in a way so that the IPTW, maximum likelihood, and doubly robust estimators are algebraically equivalent.

As our initial fit to p_0 prescribed that $\{Y|A,W\}$ followed a Gaussian distribution with fixed variance, the hardest one-dimensional submodel $\epsilon \to p_\epsilon$ for estimation of $\Psi(p_0)$ could be given by,

$$\{Y|A,W\} \sim N(Q_n^{(0)}(A,W) + \epsilon h_g(A,W), \sigma^2),$$

while the laws of $\{W\}$ and $\{A|W\}$ were left unchanged. The maximum likelihood estimator of ϵ became,

$$\epsilon_n = \frac{\sum_{i=1}^n h_g(A_i,W_i)(Y_i - Q_n^{(0)}(A_i,W_i))}{\sum_{i=1}^n h_g(A_i,W_i)},$$

leading to the updated estimate of $Q(A,W) = E[Y|A,W]$,

$$Q_n^{(1)}(A,W) = Q_n^{(0)}(A,W) + \epsilon_n h_g(A,W).$$

When the treatment mechanism g was not updated, the targeted likelihood algorithm converged in a single iteration. Note that the update did not depend in any way on the choice of variance σ^2 for the law of $\{Y|A,W\}$, so long as it was a constant. The parameter $\Psi(p_0)$ was then estimated with $\Psi(p(\epsilon_n))$, which was equal to $\Psi_{n,\text{MLE}}(Q_n^{(1)})$ and $\Psi_{n,\text{DR-IPTW}}(g, Q_n^{(1)})$. The treatment mechanism g could also be updated with targeted likelihood, to make the IPTW estimator equivalent with the maximum likelihood and double robust estimators. This was done by making a one-dimensional model $g_\epsilon(1|W)$ through $g(1|W)$ at $\epsilon = 0$, whose score at $\epsilon = 0$ was the projection of the IPTW estimator's influence curve on T_{CAR}. Such a submodel could be formed by taking,

$$\text{logit}(g_\epsilon(1|W)) = g(1|W) + \epsilon[\frac{Q(1,W)}{g(1|W)} + \frac{Q(0,W)}{g(0|W)}].$$

Because this was simply a logistic model for $\{A|W\}$, we could estimate ϵ through logistic regression. After iterating the targeted likelihood procedure to update both of the Q and g nuisance parameters until convergence, the IPTW, maximum likelihood, and double robust estimators of $\Psi(p_0)$ became equivalent.

For this data structure, $\Psi_{n,\text{DR-IPTW}}(g, Q)$ was asymptotically efficient, meaning that its asymptotic performance was superior to any other regular estimator. This efficient estimator could not be used directly on observed data, due to its dependence on the unknown nuisance paramters g and Q. We assessed the quality of an estimator Ψ_n through the ratio

$$R(\Psi_n) = \frac{E_{p_0}[n|\Psi_n - \Psi(p_0)|^2]}{E_{p_0}[n|\Psi_{n,\text{DR-IPTW}}(g, Q) - \Psi(p_0)|^2]}$$

For large enough sample size n, and consistent and asymptotically linear Ψ_n, this approximated the asymptotic relative efficiency of Ψ_n to the efficient estimator, and necessarily exceeded one. We approximated $R(\Psi_n)$ after forming Ψ_n on 1000 simulated datasets of size $n = 200$.

In our simulations, we considered known censoring mechanism g, as could occur in a randomized clinical trial. We misspecified the nuisance parameter Q, by estimating $E[Y|W]$ in the $A = 0$ and $A = 1$ strata with linear regression, while quadratic regression would have been appropriate. This first-order approximation to Q lead to an inaccurate maximum likelihood estimator, having $R(\Psi_n) = 2.63$. Confidence intervals for $R(\Psi_n)$ were negligible, due to the number of simulations. The misspecified nuisance parameter Q did not affect the performance of the IPTW estimator, or the consistency of the double robust estimator, which respectively had asymptotic relative efficiencies $R(\Psi_n)$ of 1.18 and 1.15. Note that the IPTW estimator was unbiased, but was less accurate than the double robust estimator with misspecified Q. After updating Q with a single targeted likelihood iteration, $R(\Psi_n)$ decreased to 1.10. The resulting estimator was then a maximum likelihood estimator (and double robust estimator) with updated Q, and the update greatly increased of the accuracy of the parameter estimate. When also updating the censoring mechanism g, the asymptotic relative efficiency dropped even further to 1.07, making the estimator almost equivalent with the efficient estimator. In spite of the fact that the censoring mechanism g was already known, estimating it from the data was nevertheless beneficial, as could be surmised from Chapter 2.3.7 of (van der Laan and Robins (2003)).

Thus, the targeted likelihood algorithm allowed us to estimate the nuisance parameters g and Q with maximum likelihood in a manner such that three standard estimators become identical, and led to better performance than was achieved by the initial IPTW, maximum likelihood, and double robust estimators.

6.5 Semiparametric regression example.

Let $O = (W, A, Y) \sim p_0$ and consider the semiparametric regression model $\mathcal{M} = \{p : E_p(Y \mid A, W) - E_p(Y \mid A = 0, W) = m(A, W \mid \beta(p))\}$ for some parametrization $\beta \to m(A, W \mid \beta)$ satisfying $m(0, W \mid \beta) = 0$ for all $\beta \in \mathbb{R}^d$. This is equivalent with assuming $E_0(Y \mid A, W) = m(A, W \mid \beta_0) + \theta_0(W)$ with θ_0 unspecified and $m(0, W \mid \beta) = 0$, and can therefore also be viewed as a semiparametric regression model. It has been recognized that a maximum likelihood fit (e.g., generalized additive models) of the semiparametric regression suffers from bias for the parametric part, so that one needs to undersmooth the nonparametric components in the semiparametric regression model. However, the literature does not provide practical guidance about how to undersmooth. Therefore, the targeted MLE approach presented here provides an importance practical improvement. Let $\Psi(p) = \beta(p) \in \mathbb{R}^d$ be the parameter of interest.

This type of semiparametric regression models has been considered by various authors (e.g., Newey (1995); Rosenbaum and Rubin (1983); Robins et al. (1992); Robins and Rotnitzky; Yu and van der Laan (2003)). The latter three articles derive the orthogonal complement of the nuisance tangent space (i.e., the set of all gradients of the pathwise derivative), the efficient influence curve/canonical gradient, and establish the wished double robustness of the corresponding estimating functions. In particular, for our purpose we refer to Theorem 2.1 and 2.2 in Yu and van der Laan (2003) for the following statements.

The orthogonal complement of the nuisance tangent space is given by:

$$T_{nuis}^{\perp}(p) = \{D_h(p) : h\} \subset L_0^2(P),$$

where $D_h(p)(O) \equiv (h(A, W) - E_p(h(A, W) \mid W))(Y - m(A, W \mid \beta(p)) - E_p(Y \mid A = 0, W))$. The orthogonal complement of the nuisance tangent space corresponds with the set of gradients for Ψ at p given by:

$$T_{nuis}^{\perp}(p)^* = \left\{-c(p)(h)^{-1} D_h(p)(O) : h = (h_1, \ldots, h_d)\right\},$$

where $c(p)(h) = \frac{d}{d\beta} E_p D_h(p, \beta)\Big|_{\beta=\beta(p)}$, and D_h now represents a vector function $(D_{h_1}, \ldots, D_{h_d})$. The efficient influence curve is identified by a closed form index $h(p)$ (see e.g., Yu and van der Laan (2003)), which is provided below (12). Let $D(p) = D_{h(p)}(p)$ be this efficient influence curve at p as identified by this index $h(p)$.

Let $g(p)$ be the conditional density of A, given W, under p, let $Q(p)$ be the conditional distribution of Y, given A, W, under p. We note that the

parameter $\Psi(p)$ is only a function of $Q(p)$, and the density factorizes as $p(O) = p(W)g(p)(A \mid W)Q(p)(Y \mid A, W)$. As a consequence, the elements $D_h(p)$ are orthogonal to the tangent spaces of the nuisance parameter $g(p)$ and the nuisance parameter $p(W)$. That is, we can decompose the efficient score $D(p)$ into three subcomponents as follows:

$$\begin{aligned} D(p) &= D(p) - E_p(D(p) \mid A, W) + E_p(D(p) \mid A, W) - E_p(D(p) \mid W) \\ &\quad + E_p(D(p) \mid W) - E_p D(p), \end{aligned}$$

which corresponds with scores for $p(Y \mid A, W)$, $p(A|W)$ and $p(W)$ at p, respectively, but $E_p(D(p) \mid A, W) - E_p(D(p) \mid W) = 0$ and $E_p(D(p) \mid W) - E(D(p)) = 0$. Thus the efficient influence curve $D(p)$ represents only a score for $Q(p)(Y \mid A, W)$, and indeed satisfies $E_p(D(p)(O) \mid A, W) = 0$.

Consider an initial density estimator $p_n^0 = (p_{nW}^0, g(p_n^0), Q(p_n^0))$ of (W, A, Y) with marginal distribution of W being the empirical probability distribution of $W_1, \ldots, W_n$. Above we showed that a submodel $p_n^0(\epsilon)$ through p_n^0 with score $D(p_n^0)$ at $\epsilon = 0$ can be selected to only vary the conditional density $Q(p_n^0)$ of Y, given A, W, with a score $D(p_n^0)$ at $\epsilon = 0$. Such a submodel will now be presented.

Let $p_n^0 \in \mathcal{M}$. Suppose that $Q(p_n^0)$ is a normal distribution with mean $\theta(p_n^0)(A, W) = E_{p_n^0}(Y \mid A, W)$ and variance $\sigma^2(A, W) = \sigma^2(Q_n^0)(A, W)$. Recall that $D(p_n^0) = (h(p_n^0)(A, W) - E_{p_n^0}(h(p_n^0) \mid W))(Y - m(A, W \mid \beta(p^0)) - E_{p_n^0}(Y \mid A = 0, W))$. For notational convenience, we will represent this function as $h(p_n^0)(A, W)(Y - E_{p_n^0}(Y \mid A, W))$ with now $h(p_n^0)$ so that $E_{p_n^0}(h(p_n^0)(A, W) \mid W) = 0$. Consider the parametric submodel of $\mathcal{M}$ defined as the normal density with conditional variance $\sigma^2(A, W)$ and conditional mean $m(A, W \mid \beta_n^0(\epsilon)) + \theta_n^0(\epsilon)$. That is,

$$Q_n^0(\epsilon)(Y \mid A, W) = \frac{1}{\sigma(A, W)} f_0\left(\frac{Y - m(A, W \mid \beta_n^0(\epsilon)) - \theta_n^0(\epsilon)(W)}{\sigma(A, W)}\right),$$

where $\beta_n^0(0) = \beta(Q_n^0)$, $\theta_n^0(0) = \theta(Q_n^0) = E_{Q_n^0}(Y \mid A = 0, W)$, and f_0 is the standard normal density. We note that this is a valid submodel through Q_n^0 at $\epsilon = 0$. Let $\beta(\epsilon) \equiv \beta(Q_n^0) + \epsilon$ and $\theta_n^0(\epsilon) = \theta(Q_n^0) + \epsilon^\top r$. It remains to find a function $r(W)$ so that the score of $Q_n^0(\epsilon)$ at $\epsilon = 0$ equals the efficient influence curve $D(p_n^0)$.

We have that the score $S(\epsilon)$ at ϵ is given by (note that $f_0'(x)/f_0(x) = 2x/\sigma^2$)

$$\begin{aligned} & S(\epsilon)\sigma^2(A, W) \\ &= (Y - m(A, W \mid \beta_n^0(\epsilon)) - \theta_n^0(\epsilon)(W)) \left\{ \frac{d}{d\epsilon} m(A, W \mid \beta_n^0(\epsilon)) - \frac{d}{d\epsilon} \theta_n^0(\epsilon)(W) \right\} \end{aligned}$$

$$= \left\{ \frac{d}{d\beta_n^0(\epsilon)} m(A,W \mid \beta_n^0(\epsilon)) - r(W)) \right\} (Y - m(A,W \mid \beta_n^0(\epsilon)) - \theta_n^0(\epsilon)(W)).$$

Solving for r so that $S(0) = D(p^0)$ yields the equation

$$h(p_n^0)(A,W)(Y - E_{Q^0}(Y \mid A,W)) =$$
$$\frac{1}{\sigma^2(A,W)} \left\{ \frac{d}{d\beta(Q_n^0)} m(A,W \mid \beta(Q_n^0)) - r(W) \right\} (Y - E_{Q_n^0}(Y \mid A,W)).$$

In order to have that the score equals D_h for a particular $h(A,W)$ with $E_{p_n^0}(h(A,W) \mid W) = 0$, we need

$$r(p_n^0)(W) = \frac{E_{p_n^0}\left(\frac{d/d\beta_n^0 m(A,W|\beta_n^0)}{\sigma^2(A,W)} \mid W \right)}{E_{p_n^0}\left(\frac{1}{\sigma^2(A,W)} \mid W \right)}.$$

This yields the following score for our submodel $p_n^0(\epsilon)$ at $\epsilon = 0$:

$$S(0) = h(p_n^0)(A,W)(Y - m(A,W \mid \beta(Q_n^0)) - \theta(Q_n^0)(W)),$$

where

$$\begin{aligned} h(p_n^0)(A,W) \equiv & \frac{1}{\sigma^2(A,W)} \frac{d}{d\beta(Q_n^0)} m(A,W \mid \beta(Q_n^0)) \\ & - \frac{1}{\sigma^2(A,W)} \frac{E_{p_n^0}\left(\frac{d}{d\beta(Q_n^0)} m(A,W \mid \beta(Q_n^0))/\sigma^2(A,W) \mid W \right)}{E_{p_n^0}(1/\sigma^2(A,W) \mid W)}. \end{aligned}$$

This choice $h(p_n^0)$ gives a score $S(0)$ equal to the efficient influence curve (see e.g., Yu and van der Laan (2003)). So we succeeded in finding a submodel $p_n^0(\epsilon)$ with a score at $\epsilon = 0$ equal to the efficient influence curve at p_n^0. Thus we are now ready to define the targeted MLE.

Consider the log-likelihood for $p_n^0(\epsilon)$ in ϵ:

$$l(\epsilon) \equiv \frac{1}{n} \sum_{i=1}^n \log f_0 \left(\frac{Y_i - m(A_i,W_i \mid \beta_n^0 + \epsilon) - (\theta_n^0(W) + \epsilon^\top r(p_n^0)(W))}{\sigma(A,W)} \right).$$

Let ϵ_n be the maximizer, which can thus be computed with standard weighted least squares regression:

$$\epsilon_n = \arg\min_\epsilon \sum_{i=1}^n \frac{1}{\sigma^2(A_i,W_i)} \left(Y_i - m(A_i,W_i \mid \beta_n^0 + \epsilon) - \theta_n^0(W_i) - \epsilon r(p_n^0)(W_i) \right)^2.$$

The score equation $0 = d/d\epsilon l(\epsilon) = P_n S(\epsilon)$ for ϵ_n is given by

$$0 = P_n \frac{\left\{ \frac{d}{d\beta_n^0(\epsilon)} m(\beta_n^0(\epsilon)) - r(p_n^0)) \right\} (Y - m(\beta_n^0(\epsilon)) - \theta_n^0 - \epsilon^\top r(p_n^0))}{\sigma^2}.$$

In the sequel we consider the case that $m(A,W \mid \beta) = \beta^\top m_1(A,W)$ is linear in β for some specified covariate vector $m_1(A,W)$. In this case we have $d/d\beta m(A,W \mid \beta) = m_1(A,W)$ so that the score equation $P_n S(\epsilon) = 0$ reduces to:

$$0 = P_n \frac{\{m_1 - r(p_n^0)\}\left(Y - (\beta_n^0 + \epsilon_n)m_1 - \theta_n^0 - \epsilon_n^\top r(p_n^0)\right)}{\sigma^2}. \tag{12}$$

Firstly, we note that ϵ_n exist in closed form:

$$\epsilon_n = A_n^{-1} P_n \frac{\{m_1 - r(p_n^0)\}\left(Y - \beta_n^{0\top} m_1 - \theta_n^0\right)}{\sigma^2},$$

where the $d \times d$ matrix A_n is given by

$$A_n \equiv \frac{1}{n}\sum_{i=1}^n \frac{1}{\sigma^2(A_i,W_i)} \left\{m_1(A_i,W_i) - r(p_n^0)(W_i)\right\}\left(m_1(A_i,W_i) + r(p_n^0)(W_i)\right)^\top.$$

Let $p_n^0(\epsilon_n)$ be the new density estimator. Recall that the distribution of (A,W) under $p_n^0(\epsilon_n)$ is still the same as under p_n^0, because $p_n^0(\epsilon)$ only updates the conditional distribution of Y, given A,W. We now wish to investigate if the first step targeted MLE $p_n^1 \equiv p_n^0(\epsilon_n)$ already solves the efficient score equation: $P_n D(p_n^1) = P_n D(p_n^0(\epsilon_n)) = 0$. We have that $P_n D(p_n^0(\epsilon_n))$ is given by

$$P_n \frac{\{m_1 - r(p_n^0(\epsilon_n))\}\left(Y - (\beta_n^0 + \epsilon_n)m_1 - \theta_n^0 - \epsilon_n r(p^0(\epsilon_n))\right)}{\sigma^2}.$$

Because $r(p_n^0(\epsilon)) = r(p_n^0)$, it follows that $P_n D(p^0(\epsilon_n))$ is given by

$$P_n \frac{\{m_1 - r(p_n^0)\}\left(Y - (\beta_n^0 + \epsilon_n)m_1 - \theta_n^0 - \epsilon_n r(p_n^0)\right)}{\sigma^2},$$

but the latter equals zero by the fact that $P_n S(\epsilon_n) = 0$ (12). This proves that, if $m(A,W \mid \beta)$ is linear in β, then the targeted maximum likelihood estimator is achieved in the first step of the algorithm and solves the efficient influence curve estimating equation $P_n D(p) = 0$. If one would also update $\sigma^2(A,W)$ in the submodel $p_n^0(\epsilon)$, then the algorithm would have to be iterated in order to converge to a targeted MLE solving $P_n D(p) = 0$. Similarly, for nonlinear models $m(A,W \mid \beta)$ the targeted MLE algorithm will also need to be iterated till convergence.

7 Targeted MLE as loss based estimation.

In the previous sections we defined a targeted MLE in terms of an initial density estimator and the targeted MLE algorithm applied to this initial density estimator. In order to provide a general data adaptive likelihood based

approach for construction of targeted MLE's (also allowing for an integrated data adaptive approach for searching over the initial densities, just as in sieve based MLE), we now note that the targeted MLE approach corresponds with a particular modified log-likelihood loss function. Specifically, let

$$L(p \mid P_0) \equiv -\log p^*(p),$$

where $p^*(p)$ is defined as the limit for $k \to \infty$ of the targeted MLE applied to P_0 and starting at p:

$$p^{k+1} = arg \max_{p \in \{p^k(\epsilon):\epsilon\}} P_0 \log p. \tag{13}$$

Note that $L(p \mid P_0)$ is a loss function for densities p of the data indexed by unknown nuisance parameters, since the $\epsilon_0^k \equiv \arg\max_\epsilon P_0 \log p^k(\epsilon)$ are unknown. However, estimation of the unknown nuisance parameter corresponds simply with applying the targeted MLE algorithm to the data starting at p. The loss function satisfies

$$p_0 = \arg\min_{p \in \mathcal{M}} P_0 L(p \mid P_0),$$

because $p^*(p_0) = p_0$ and $p_0 = \arg\min_{p \in \mathcal{M}} -P_0 \log p$. Therefore, we can apply the unified loss based learning approach presented in van der Laan and Dudoit (2003) based on this new loss function $L(p \mid P_0)$ for a candidate density p. Succinctly, this loss based learning approach works as follows. Let $\mathcal{M}_s \subset \mathcal{M}$ be a sieve of $\mathcal{M}$ indexed by fine tuning parameters s. Let

$$p_{sn} = \hat{\Phi}_s(P_n) \equiv \arg\min_{p \in \mathcal{M}_s} P_n L(p \mid P_n) = \arg\max_{p \in \mathcal{M}_s} P_n \log p_n^*(p),$$

where $p_n^*(p)$ represents the limit density of the targeted MLE algorithm starting at p applied to the data P_n. Note that this maximization corresponds with maximizing the log likelihood over solutions of $P_n D(p^*) = 0$, where the $p^* = p^*(p)$ is restricted by the constraints on the initial p. We can select s with likelihood based cross-validation:

$$s_n = \hat{S}(P_n) \equiv \arg\min_s E_{B_n} P_{n,B_n}^1 L(\hat{\Phi}_s(P_{n,B_n}^0) \mid P_{n,B_n}^0),$$

resulting in the targeted ML density estimator

$$p_n \equiv p_{s_n n} = \hat{\Phi}_{\hat{S}(P_n)}(P_n)$$

and targeted ML estimator of ψ_0 given by $\psi_n = \Psi(p_n)$.

8 Discussion.

In this article we assumed a model in terms of densities with respect to a known dominating measure, and our targeted MLE density estimators are assumed to be dominated by this dominating measure. This allowed us to simplify the presentation of the method. However, we also wish to stress that the presented targeted maximum likelihood estimation methodology can easily be generalized to targeted maximum likelihood estimation in models in terms of probability distributions including (say) discrete as well as continuous distributions, just as this is common practice in maximum likelihood estimation in semiparametric models. The targeted MLE algorithm takes as input an initial density with respect to a specified dominating measure, and is based on a hardest submodel in terms of densities with respect to this same dominating measure. Thus, the targeted MLE algorithm can be applied to discrete distributions as well as continuous distributions, and as a consequence, the (loss based) targeted MLE learning as presented in Section 7 applies to models that are not necessarily dominated by a single dominating measure.

As a further generalization, the iterative principle underlying this work can be applied to loss functions other than the negative log likelihood. Given a loss function defined on the data and parameter space (and possibly a nuisance parameter η), we can make a one-dimensional ϵ-extension through a space containing both the parameter Ψ and nuisance parameter η, initialize the parameter estimate at $\Psi(0)$, and then update the parameter estimate by choosing ϵ to minimize the empirical risk $\frac{1}{n}\sum_{i=1}^{n} L(O_i, \Psi(\epsilon)|\eta(\epsilon))$. The requirement underlying the procedure is that $\frac{d}{d\epsilon}L(O, \Psi(\epsilon)|\eta(\epsilon))|_{\epsilon=0}$ is equal to an estimating equation for the parameter Ψ. If this condition is met, then solving this estimating equation should correspond to convergence of the iterative empirical risk minimization algorithm. Hence, applying the algorithm with such a loss function $L(O, \Psi|\eta)$ leads to a fusion of general loss based estimation and estimating function methodology.

Given a density estimator we defined a targeted density estimator through an iterative maximum likelihood algorithm along hardest submodels with a score equal to the efficient influence curve of the parameter of interest. This tool allows us to map any candidate density p into its targeted version $p_n^*(p)$. We now showed that by using the minus log density as loss function and thereby use the log-likelihood criteria in combination with the cross-validated log-likelihood criteria, *but restricted to targeted density estimators only*, we can build data adaptive sieve based algorithms for generating a final targeted ML density estimator and corresponding substitution estimator of the parameter of interest.

By restricting the log-likelihood criteria and cross-validated log-likelihood criteria to targeted densities only, targeted maximum likelihood estimation provides now a purely likelihood based methodology for estimation of any kind of parameter such as pathwise differentiable parameters and infinite dimensional parameters: see our accompanying technical report.

In particular, we showed that targeted maximum likelihood estimation completely unifies maximum likelihood estimation and estimating function based estimation, and results in important improvements in both. Targeted MLE also deals naturally with the issue of multiple solutions of estimating equations by using the log-likelihood as the criteria to be maximized. Another nice feature of targeted MLE is that it always improves on the initial density estimator by increasing the log-likelihood fit. As a consequence, when targeted MLE is applied to estimate pathwise differentiable parameters of a full data distribution F_X in CAR censored data models as in (van der Laan and Robins (2003)), if one applies the targeted MLE to an initial $p_n^0 = (g_n^0, Q_n^0)$ with g_n^0 and Q_n^0 being fits of the censoring mechanism g_0 and the F_X-factor Q_0 of the density p_0, then it provides an estimator which is guaranteed to be more efficient than the double robust IPCW estimator based on estimating the nuisance parameters (g_0, Q_0) with p_n^0. So the targeted MLE algorithm provides a natural way to always improve on any initial double robust IPCW locally efficient estimator as presented in van der Laan and Robins (2003).

References

P.J. Bickel, A.J. Klaassen, Y. Ritov, and J.A. Wellner. *Efficient and adaptive inference in semiparametric models*. Johns Hopkins university press, Baltimore, 1993a.

P.J. Bickel, C.A.J. Klaassen, Y. Ritov, and J. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, 1993b.

S.R. Cosslett. Efficient semiparametric estimation of censored and truncated regressions via smooth self-consistency equation. *Econometrica*, 72(4):1277–1284, 2004.

R.D. Gill, M.J. van der Laan, and J.M. Robins. Coarsening at random: characterizations, conjectures and counter-examples. In D.Y. Lin and T.R. Fleming, editors, *Proceedings of the First Seattle Symposium in Biostatistics*, pages 255–94, New York, 1997. Springer Verlag.

D.F. Heitjan and D.B. Rubin. Ignorability and coarse data. *Annals of statistics*, 19(4):2244–2253, December 1991.

M. Jacobsen and N. Keiding. Coarsening at random in general sample spaces and random censoring in continuous time. *Annals of Statistics*, 23:774–86, 1995.

C.A.J. Klaassen. Consistent estimation of the influence function of locally asymptotically linear estimators. *Annals of Statistics*, 15:1548–1562, 1987.

W.K. Newey. Semiparametric efficiency bounds. *Journal of applied econometrics*, 1(4):335–341, 1995. ISSN 1350-7265.

J. M. Robins and A. Rotnitzky. Comment on the Bickel and Kwon article, "Inference for semiparametric models: Some questions and an answer". *Statistica Sinica*, 11(4):920–936, 2001.

J. M. Robins, A. Rotnitzky, and M.J. van der Laan. Comment on "On Profile Likelihood" by S.A. Murphy and A.W. van der Vaart. *Journal of the American Statistical Association – Theory and Methods*, 450:431–435, 2000.

J.M. Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, 2000.

J.M. Robins, S.D Mark, and W.K. Newey. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48:479–495, 1992.

J.M Robins and A. Rotnitzky. Comment on Inference for semiparametric models: some questions and an answer, by Bickel, P.J. and Kwon.

J.M. Robins and A. Rotnitzky. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology*, Methodological issues. Bikhäuser, 1992.

P.R. Rosenbaum and D.B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

M.J. van der Laan. *Efficient and Inefficient Estimation in Semiparametric Models*. Centre of Mathematics and Computer Science (CWI), Amsterdam, 1995.

M.J. van der Laan. Identity for npmle in censored data models. *Lifetime Data Models*, 4(0):83–102, 1998.

M.J. van der Laan. Statistical inference for variable importance. *International Journal of Biostatistics*, 2(1), 2006.

M.J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Technical report, Division of Biostatistics, University of California, Berkeley, November 2003.

M.J. van der Laan and J.M. Robins. *Unified methods for censored longitudinal data and causality*. Springer, New York, 2003.

M.J. van der Laan and D. Rubin. Estimating function based cross-validation and learning. Technical report 180, Division of Biostatistics, University of California, Berkeley, 2005.

M.J. van der Laan and D. Rubin. Estimating function based cross-validation. In J. Fan and H.L. Koul, editors, *Frontiers of Statistics*, pages 87–108. Imperial College Press, 2006.

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, 1996a.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Emprical Processes*. Springer-Verlag New York, 1996b.

Z. Yu and M.J. van der Laan. Measuring treatment effects using semiparametric models. Technical report, Division of Biostatistics, University of California, Berkeley, 2003.

Chapter 3

Super (Machine) Learning using Cross Validation

3.1 *Super Learner*

The following article appears as it was originally published on the University of California, Berkeley Division of Biostatistics Working Paper Series website in 2007, http://www.bepress.com/ucbbiostat/paper222/.

It was later published in *Statistical Applications in Genetics and Molecular Biology* in 2007, http://www.bepress.com/sagmb/vol6/iss1/art25/.

Super Learning

Mark J. van der Laan, Eric C. Polley and Alan E. Hubbard
Division of Biostatistics, University of California, Berkeley
laan@stat.berkeley.edu

Abstract

Previous articles (van der Laan and Dudoit (2003); van der Laan et al. (2006); Sinisi et al. (2007)) advertised and theoretically validated the use of cross-validation to select among many candidate estimators to compute a so called super learner which outperforms any of the given candidate estimators. The theoretical basis was provided for this super learner based on oracle results for the cross-validation selector (e.g., van der Laan and Dudoit (2003); van der Laan et al. (2006)) and in Sinisi et al. (2007). In addition, these papers contained a practical demonstration of the adaptivity of this so called super learner in the context of prediction of the fitness of the HIV virus as a function of its mutations. This article proposes a fast algorithm for constructing a super learner in prediction which uses V-fold cross-validation to select a functional form of an initial set of candidate predictors according to a parametric or semi-parametric model, or possibly, data adaptively. The paper contains a proof that the resulting super learner performs asymptotically as well as the oracle selector among the continuum of estimators defined by the (semi-)parametric functional forms of the initial set of candidate estimators.

This approach also yields a new class of cross-validation methods to select among a family of candidate estimators by formulating the minimization of the cross-validated risk over the family of candidate estimators as a new least squares regression problem which itself can be carried out with any type of parametric or nonparametric regression methodology (e.g. using cross-validation itself), thereby preventing over-fitting of the cross-validated risk. Simulations and data analysis suggest this new proposed super learner superior to competing methods. This approach for construction of a super learner generalizes to any parameter which can be defined as a minimizer of a loss function.

1 Introduction

Numerous methods exist to learn from data the best predictor of a given outcome based on a sample of n independent and identically distributed observations $O_i = (Y_i, X_i)$, Y_i the outcome of interest, and X_i a vector of input variables, $i = 1, \ldots, n$. A few examples include decision trees, neural networks, support vector regression, least angle regression, logic regression, poly-class, Multivariate Adaptive Regression Splines (MARS), and the Deletion/Substitution/Addition (D/S/A) algorithm. Such learners can be characterized by the mechanism used to search the parameter space of possible regression functions. For example, the D/S/A algorithm (Sinisi and van der Laan, 2004) uses polynomial basis functions, while logic regression (Ruczinski et al., 2003) constructs Boolean expressions of binary covariates. The performance of a particular learner depends on how effective its searching strategy is in approximating the optimal predictor defined by the true data generating distribution. Thus, the relative performance of various learners will depend on the true data-generating distribution. In practice, it is generally impossible to know *a priori* which learner will perform best for a given prediction problem and data set. To solve the problem, some researchers have proposed combining learners in various methods and have exhibited better performance over a single candidate learner (Freund et al., 1997; Hansen, 1998), but there is concern that these methods may over-fit the data and may not be the optimal way to combine the candidate learners.

The framework for unified loss-based estimation (van der Laan and Dudoit, 2003) suggests a solution to this problem in the form of a new learner, termed the "super learner". In the context of prediction, this learner is itself a prediction algorithm, which applies a set of candidate learners to the observed data, and chooses the optimal learner for a given prediction problem based on cross-validated risk. Theoretical results show that such a super learner will perform asymptotically as well as or better than any of the candidate learners (van der Laan and Dudoit, 2003; van der Laan et al., 2006).

To be specific, consider some candidate learners. *Least Angle Regression* (LARS) (Efron et al., 2004) is a model selection algorithm related to the lasso. *Logic Regression* (Ruczinski et al., 2003) is an adaptive regression methodology that attempts to construct predictors as Boolean combinations of binary covariates. The D/S/A algorithm (Sinisi and van der Laan, 2004) for polynomial regression data-adaptively generates candidate predictors as polynomial combinations of continuous and/or binary covariates, and is avail-

Method	R Package	Authors
Least Angle Regression	lars	Hastie and Efron
Logic Regression	LogicReg	Kooperberg and Ruczinski
D/S/A	DSA	Neugebauer and Bullard
Regression Trees	rpart	Therneau and Atkinson
Ridge Regression	MASS	Venables and Ripley
Random Foretss	randomForest	Liaw and Wiener
Adaptive Regression Splines	polspline	Kooperberg

Table 1: R Packages for Candidate Learners. R is available at http://www.r-project.org

able as an `R` package at http://www.stat.berkeley.edu/users/laan/Software/. *Classification and Regression Trees* (CART) (Breiman et al., 1984) builds a recursive partition of the covariates. Another candidate learner is random forests Breiman (2001), which is a random bootstrap version of the regression tree. *Ridge Regression* (Hoerl and Kennard, 1970) minimizes a penalized least squares with a penalty on the $L2$ norm of the parameter vector. *Multivariate Adaptive Regression Splines* (MARS) Friedman (1991) is an automated model selection algorithm which creates a regression spline function. Table 1 contains citations of R packages for each of the candidate learners. All of these methods have the option to carry out selection using v-fold cross-validation. The selected fine-tuning parameter(s) can include the ratio of the $L1$ norm of the coefficient vector in LARS to the norm of the coefficient vector from least squares; the number of logic trees and leaves in Logic Regression; and the number of terms and a complexity measure on each of the terms in D/S/A.

Cross-validation divides the available *learning* set into a *training* set and a *validation* set. Observations in the training set are used to construct (or *train*) the learners, and observations in the validation set are used to assess the performance of (or *validate*) these learners. The cross-validation selector selects the learner with the best performance on the validation sets. In v-fold cross-validation, the learning set is divided into v mutually exclusive and exhaustive sets of as nearly equal size as possible. Each set and its complement play the role of the validation and training sample, respectively, giving v splits of the learning sample into a training and corresponding validation sample. For each of the v splits, the estimator is applied to the training

set, and its risk is estimated with the corresponding validation set. For each learner the v risks over the v validation sets are averaged resulting in the so-called *cross-validated risk*. The learner with the minimal cross-validated risk is selected.

It is helpful to consider each learner as an algorithm applied to empirical distributions. Thus, if we index a particular learner with an index k, then this learner can be represented as a function $P_n \rightarrow \hat{\Psi}_k(P_n)$ from empirical probability distributions P_n to functions of the covariates. Consider a collection of $K(n)$ learners $\hat{\Psi}_k$, $k = 1, \ldots, K(n)$, in parameter space $\mathbf{\Psi}$. The super learner is a new learner defined as

$$\hat{\Psi}(P_n) \equiv \hat{\Psi}_{\hat{K}(P_n)}(P_n),$$

where $\hat{K}(P_n)$ denotes the cross-validation selector described above which simply selects the learner which performed best in terms of cross-validated risk. Specifically,

$$\hat{K}(P_n) \equiv \arg\min_k E_{B_n} \sum_{i, B_n(i)=1} (Y_i - \hat{\Psi}_k(P^0_{n,B_n})(X_i))^2,$$

where $B_n \in \{0,1\}^n$ denotes a random binary vector whose realizations define a split of the learning sample into a training sample $\{i : B_n(i) = 0\}$ and validation sample $\{i : B_n(i) = 1\}$. Here P^1_{n,B_n} and P^0_{n,B_n} are the empirical probability distributions of the validation and training sample, respectively.

The aggressive use of cross-validation is inspired by the theorem 3.1 in van der Laan et al. (2006). The theorem is provided in the appendix.

The "oracle" selector is defined in Theorem 2 in the appendix as the estimator, among the $K(n)$ learners considered, which minimizes risk under the true data-generating distribution. In other words, the oracle selector is the best possible estimator given the set of candidate learners considered; however, it depends on both the observed data and P_0, and thus is unknown.

This theorem shows us that the super learner performs as well (in terms of expected risk difference) as the oracle selector, up to a typically second order term. Thus, as long as the number of candidate learners considered ($K(n)$) is polynomial in sample size, the super learner is the optimal learner in the following sense:

- If, as is typical, none of the candidate learners (nor, as a result, the oracle selector) converge at a parametric rate, the super learner performs asymptotically as well (in the risk difference sense) as the oracle selector, which chooses the best of the candidate learners.

- If one of the candidate learners searches within a parametric model and that parametric model contains the truth, and thus achieves a parametric rate of convergence, then the super learner achieves the almost parametric rate of convergence $\log n/n$.

Organization: The current article builds and extends this super learning methodology. In section 2 we will describe our new proposal for super learning, also using an initial set of candidate learners and cross-validation as above, but now allowing for semi-parametric families of the candidate learners, and formulating the minimization of cross-validated risk as another regression problem for which one can select an appropriate regression methodology (e.g involving cross-validation or penalized regression). This is an important improvement relative to our previous super learning proposal by 1) extending the set of initial candidate learners into a large family of candidate learners one obtains by combining the initial candidate learners according to a parametric or semi-parametric model, thereby obtain a potentially much more flexible learner, and 2) by controlling over-fitting of the cross-validated risk through the use of data adaptive regression algorithms using cross-validation or penalization itself. Importantly, these gains come at no cost regarding computing time. In Section 3 we investigate the practical performance of this new super learning algorithm based on simulated as well as a number of real data sets.

2 The proposed super learning algorithm

Suppose one observes n i.i.d. observations $O_i = (X_i, Y_i) \sim P_0$, $i = 1, \ldots, n$, and the goal is to estimate the regression $\psi_0(X) = E_0(Y \mid X)$ of $Y \in \mathcal{Y}$ on $X \in \mathcal{X}$. The regression can be defined as the minimizer of the expectation of the squared error loss function:

$$\psi_0 = \arg\min_{\psi} E_0 L(O, \psi),$$

where $L(O, \psi) = (Y - \psi(X))^2$. The proposed super learner immediately applies to any parameters that can be defined as minimizers of a loss function $L(O, \psi)$ over a parameter space $\mathbf{\Psi}$, but the article focuses on the prediction problem using the squared error loss function.

Let $\hat{\Psi}_j$, $j = 1, \ldots, J$, be a collection of J candidate learners, which represent mappings from the empirical probability distribution P_n into the parameter space $\mathbf{\Psi}$ consisting of functions of X.

The proposed super learner uses V-fold cross-validation. Let $v \in \{1, \ldots, V\}$ index a sample split into a validation sample $V(v) \subset \{1, \ldots, n\}$ and training sample (the complement of $V(v)$) $T(v) \subset \{1, \ldots, n\}$, where $V(v) \cup T(v) = \{1, \ldots, n\}$. Here we note that the union, $\cup_{v=1}^{V} V(v) = \{1, \ldots, n\}$, of the validation samples equals the total sample, and the validations samples are disjoint: $V(v_1) \cap V(v_2) = \emptyset$ for $v_1 \neq v_2$. For each $v \in \{1, \ldots, V\}$, let, $\psi_{njv} \equiv \hat{\Psi}_j(P_{nT(v)})$ be the realization of the j^{th}-estimator $\hat{\Psi}_j$ when applied to the training sample $P_{nT(v)}$.

For an observation i, let $v(i)$ denote the validation sample it belongs to, $i = 1, \ldots, n$. We now construct a new data set of n observations as follows: (Y_i, Z_i), where $Z_i \equiv (\psi_{njv(i)}(X_i) : j = 1, \ldots, J)$ is the vector consisting of the J predicted values according to the J estimators trained on the training sample $P_{nT(v(i))}$, $i = 1, \ldots, n$. Let $\mathcal{Z}$ be the set of possible outcomes for Z.

Minimum cross-validated risk predictor: Another input of this super learning algorithm is yet another user-supplied prediction algorithm $\tilde{\Psi}$ that estimates the regression $E(Y \mid Z)$ of Y onto Z based on the data set (Y_i, Z_i), $i = 1, \ldots, n$. For notational convenience, we will denote $\{(Y_i, Z_i) : i = 1, \ldots, n\}$ with $P_{n,Y,Z}$, so that $\tilde{\Psi}$ is a mapping from $P_{n,Y,Z}$ to $\tilde{\Psi}(P_{n,Y,Z}) : \mathcal{Z} \to \mathcal{Y}$, where the latter is a function from $\mathcal{Z}$ to $\mathcal{Y}$. We will refer to this algorithm $\tilde{\Psi}$ as the minimum cross-validated risk predictor since it aims to minimize the cross-validated risk, $\tilde{\psi} \to \sum_{i=1}^{n}(Y_i - \tilde{\psi}(Z_i))^2$, over a set of candidate functions $\tilde{\psi}$ from $\mathcal{Z}$ into $\mathcal{Y}$, although, we allow penalization or cross-validation to avoid over-fitting of this cross-validated risk criteria.

This now defines a mapping $\hat{\Psi}^*$ from the original data $P_n \equiv \{Y_i, X_i) : i = 1, \ldots, n\}$ into the predictor

$$\tilde{\Psi}\left(\{Y_i, Z_i = (\hat{\Psi}_j(P_{nT(v_i)})(X_i) : j = 1, \ldots, J)) : i = 1, \ldots, n\}\right)$$

obtained by applying the cross-validated risk minimizer $\tilde{\Psi}$ to $P_{n,Y,Z} = \{(Y_i, Z_i) : i = 1, \ldots, n\}$. Denote $\psi_n^* = \hat{\Psi}^*(P_n)$ as the actual obtained predictor when one applies the learner $\hat{\Psi}^*$ to the original sample P_n. We note that $\psi_n^* \in \boldsymbol{\Psi}^* \equiv \{f : \mathcal{Z} \to \mathcal{Y}\}$ is a function of Z into the outcome set $\mathcal{Y}$ for Y.

The super learner for a value X based on the data (i.e., P_n) is now given by

$$\hat{\Psi}(P_n)(X) \equiv \hat{\Psi}^*(P_n)((\hat{\Psi}_j(P_n)(X), j = 1, \ldots, J). \tag{1}$$

In words, the super learner of Y for a value X is obtained by evaluating the predictor $\psi_n^* = \hat{\Psi}^*(P_n)$ at the J predicted values, $\hat{\Psi}_j(P_n)(X)$, at X of the J

candidate learners. Figure 1 contains a flow diagram for the steps involved in the super learner.

2.1 Specific choices of the minimum cross-validated risk predictor.

Parametric minimum cross-validated risk predictor: Consider a few concrete choices that aim to fit a regression of Y onto the J predicted values Z based on the corresponding training samples from (Y_i, Z_i), $i = 1, \ldots, n$ for the algorithm $\hat{\Psi}^*$. Define the cross-validated risk criteria:

$$R_{CV}(\beta) \equiv \sum_{i=1}^{n} (Y_i - m(Z_i \mid \beta))^2,$$

where one could use, for example, the linear regression model $m(z \mid \beta) = \beta z$. If $Y \in \{0, 1\}$, then one could use the logistic linear regression model $m(z \mid \beta) = 1/(1 + \exp(-\beta z))$, if one allows predictions in the range of $[0, 1]$, or, if one wants a predictor mapping into $\{0, 1\}$, then we can choose $m(z \mid \alpha_0, \beta) \equiv I(1/(1 + \exp(-\beta z)) > \alpha_0)$ as the indicator that the logistic regression score exceeds a cut-off α_0. Let $\beta_n = \arg\min_\beta R_{CV}(\beta)$ be the least squares or MLE estimator, and let

$$\psi_n^*(z) \equiv m(z \mid \beta_n).$$

One could also estimate β with a constrained least squares regression estimator such as penalized L_1-regression (Lasso), penalized L_2 regression (shrinkage), where the constraints are selected with cross-validation, or one could restrict β to the set of positive weights summing up till 1.

Data adaptive minimum cross-validated risk predictor: There is no need to restrict ψ_n^* to parametric regression fits. For example, one could define ψ_n^* in terms of the application of a particular data adaptive (machine learning) regression algorithm to the data set (Y_i, Z_i), $i = 1, \ldots, n$, such as CART, D/S/A, or MARS, among others. In fact, one could apply a super learning algorithm itself to estimate $E(Y \mid Z)$. In this manner one can let the data speak in order to build a good predictor of Y based on covariate vector Z based on (Y_i, Z_i), $i = 1, \ldots, n$.

Thus, this super learner is indexed, beyond the choice of initial candidate estimators, by a choice of minimum cross-validated risk predictor. As a consequence, the proposal provides a whole class of tools indexed by an arbitrary

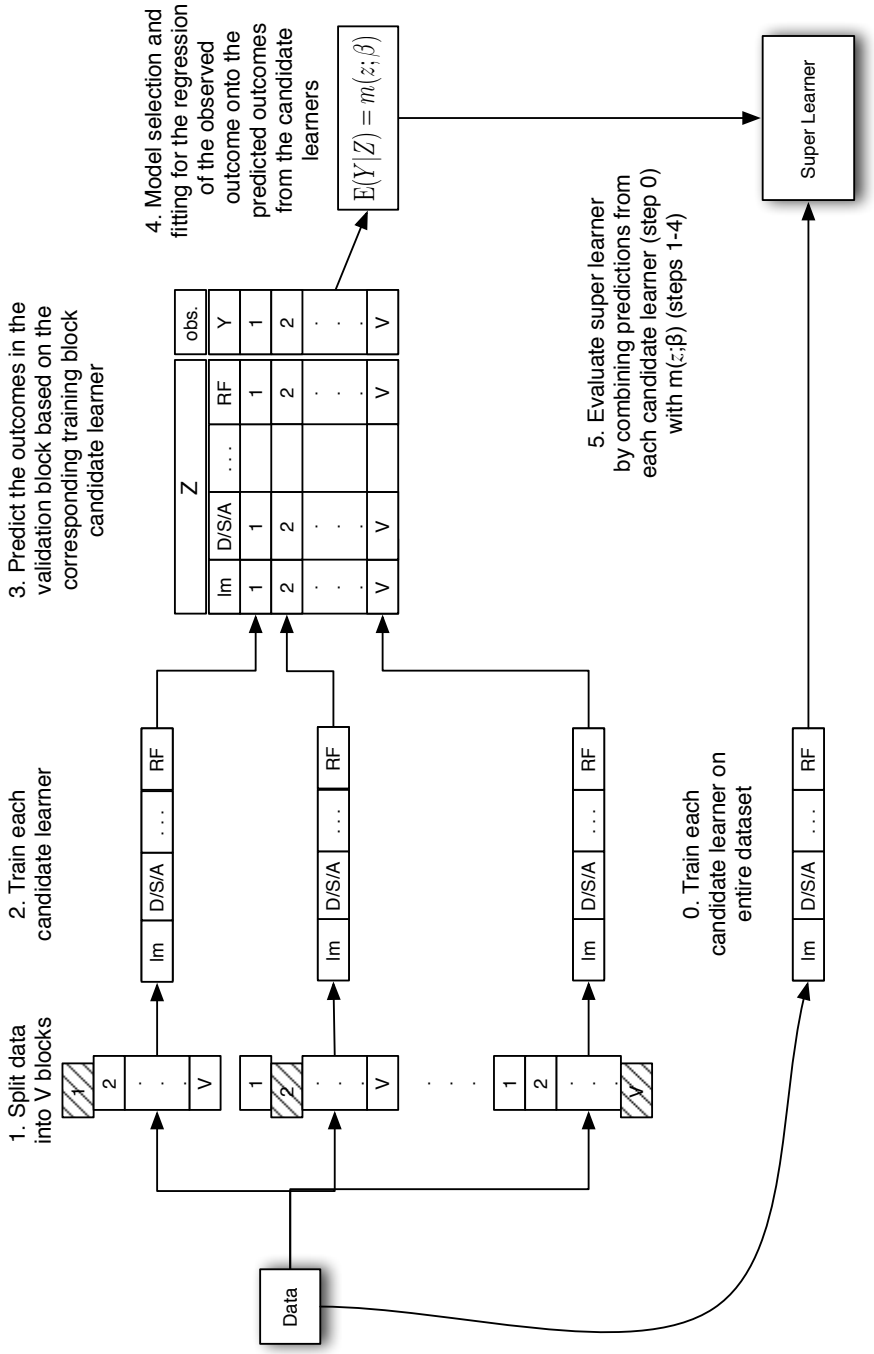


Figure 1: Flow Diagram for Super Learner

choice of regression algorithm (i.e., ψ_n^*) to map a set of candidate learners into a new cross-validated estimator (i.e. super learner). In particular, it provides a new way of using the cross-validated risk function, which goes beyond minimizing the cross-validated risk over a set of candidate learners.

3 Finite sample result and asymptotics for the super learner.

An immediate consequence of Theorem 2 above is the following result for the proposed super learner (1), which provides for the case that the minimum cross-validated risk predictor is based on a parametric regression model.

Theorem 1 *Assume $P((Y,X) \in \mathcal{Y} \times \mathcal{X}) = 1$, where $\mathcal{Y}$ is a bounded set in $\mathbb{R}$, and $\mathcal{X}$ is a bounded Euclidean set. Assume that the candidate estimators map into $\mathcal{Y}$: $P(\hat{\Psi}_j(P_n) \in \mathcal{Y}, j = 1, \ldots, J) = 1$.*

Let $v \in \{1, \ldots, V\}$ index a sample split into a validation sample $V(v) \subset \{1, \ldots, n\}$ and corresponding training sample $T(v) \subset \{1, \ldots, n\}$ (complement of $V(v)$), where $V(v) \cup T(v) = \{1, \ldots, n\}$, and $\cup_{v=1}^{V} V(v) = \{1, \ldots, n\}$. For each $v \in \{1, \ldots, V\}$, let, $\psi_{njv} \equiv \hat{\Psi}_j(P_{nT(v)})$, $\mathcal{X} \to \mathcal{Y}$, be the realization of the j-th estimator $\hat{\Psi}_j$ when applied to the training sample $T(v)$.

For an observation i let $v(i)$ be the validation sample observation i belongs to, $i = 1, \ldots, n$. Construct a new data set of n observations defined as: (Y_i, Z_i), where $Z_i \equiv (\psi_{njv(i)}(X_i) : j = 1, \ldots, J) \in \mathcal{Y}^J$ is the J-dimensional vector consisting of the J predicted values according to the J estimators trained on the training sample $T(v(i))$, $i = 1, \ldots, n$.

Consider a regression model $z \to m(z \mid \alpha)$ for $E(Y \mid Z)$ indexed by a $\alpha \in \mathcal{A}$ representing a set of functions from $\mathcal{Y}^J$ into $\mathcal{Y}$. Consider a grid (or any finite subset) $\mathcal{A}_n$ of α-values in the parameter space $\mathcal{A}$. Let $K(n) =\mid \mathcal{A}_n \mid$ be the number of grid points which grows at most at a polynomial rate in n: $K(n) \leq n^q$ for some $q < \infty$.

Let

$$\alpha_n \equiv \arg\min_{\alpha \in \mathcal{A}_n} \sum_{i=1}^{n} (Y_i - m(Z_i \mid \alpha))^2.$$

Consider the regression estimator $\psi_n : \mathcal{X} \to \mathcal{Y}$ defined as

$$\psi_n(x) \equiv m((\psi_{jn}(x) : j = 1, \ldots, J) \mid \alpha_n).$$

For each $\alpha \in \mathcal{A}$, define the candidate estimator $\hat{\Psi}_\alpha(P_n) \equiv m((\hat{\Psi}_j(P_n) : j = 1, \ldots, J) \mid \alpha)$: i.e.

$$\hat{\Psi}_\alpha(P_n)(x) = m((\hat{\Psi}_j(P_n)(x) : j = 1, \ldots, J) \mid \alpha).$$

Consider the oracle selector of α:

$$\tilde{\alpha}_n \equiv \arg\min_{\alpha \in \mathcal{A}_n} \frac{1}{V} \sum_{v=1}^{V} d(\hat{\Psi}_\alpha(P_{nT(v)}), \psi_0),$$

where

$$d(\psi, \psi_0) = E_0(L(X, \psi) - L(X, \psi_0)) = E_0(\psi(X) - \psi_0(X))^2.$$

For each $\delta > 0$ we have that there exists a $C(\delta) < \infty$ such that

$$\frac{1}{V} \sum_{v=1}^{V} E d(\hat{\Psi}_{\alpha_n}(P_{nT(v)}), \psi_0) \leq (1+\delta) E \min_{\alpha \in \mathcal{A}_n} \frac{1}{V} \sum_{v=1}^{V} d(\hat{\Psi}_\alpha(P_{nT(v)}), \psi_0) + C(\delta) \frac{V \log n}{n}.$$

Thus, if

$$\frac{E \min_{\alpha \in \mathcal{A}_n} \frac{1}{V} \sum_{v=1}^{V} d(\hat{\Psi}_\alpha(P_{nT(v)}), \psi_0)}{\frac{\log n}{n}} \to 0 \text{ as } n \to \infty, \tag{2}$$

then it follows that the estimator $\hat{\Psi}_{\alpha_n}$ is asymptotically equivalent with the oracle estimator $\hat{\Psi}_{\tilde{\alpha}_n}$ when applied to samples of size $(1 - 1/V)n$:

$$\frac{\frac{1}{V} \sum_{v=1}^{V} E d(\hat{\Psi}_{\alpha_n}(P_{nT(v)}), \psi_0)}{E \min_{\alpha \in \mathcal{A}_n} \frac{1}{V} \sum_{v=1}^{V} d(\hat{\Psi}_\alpha(P_{nT(v)}), \psi_0)} \to 1 \text{ as } n \to \infty.$$

If (2) does not hold, then it follows that $\hat{\Psi}_{\alpha_n}$ achieves the $(\log n)/n$ rate:

$$\frac{1}{V} \sum_{v=1}^{V} E d(\hat{\Psi}_{\alpha_n}(P_{nT(v)}), \psi_0) = O\left(\frac{\log n}{n}\right).$$

Discussion of conditions. The discrete approximation $\mathcal{A}_n$ of $\mathcal{A}$ used in this theorem is typically asymptotically negligible. For example, if $\mathcal{A}$ is a bounded Euclidean set, then the distance between neighboring points on the grid can be chosen as small as $1/n^q$ for some $q < \infty$ so that minimizing

a criteria over such a fine grid $\mathcal{A}_n$ versus minimizing over the whole set $\mathcal{A}$ results in asymptotically equivalent procedures. For example, if α is a Euclidean parameter and $\| m(\cdot \mid \alpha_1) - m(\cdot \mid \alpha_2) \|_\infty < C \| \alpha_1 - \alpha_2 \|$ for some $C < \infty$, where $\| \cdot \|_\infty$ denotes the supremum norm, then it follows that for each $\delta > 0$ we have that there exists a $C(\delta) < \infty$ such that

$$\frac{1}{V}\sum_{v=1}^{V} E d(\hat{\Psi}_{\alpha_n}(P_{nT(v)}), \psi_0) \leq (1+\delta) E \min_{\alpha \in \mathcal{A}} \frac{1}{V}\sum_{v=1}^{V} d(\hat{\Psi}_\alpha(P_{nT(v)}), \psi_0) + C(\delta)\frac{\log n}{n},$$

where $\alpha_n = \arg\min_{\alpha \in \mathcal{A}} \sum_{i=1}^{n} (Y_i - m(Z_i \mid \alpha))^2$. The other conclusions of the theorem now also apply.

This theorem implies that the selected prediction algorithm $\hat{\Psi}_{\alpha_n}$ will either perform asymptotically as well (up till the constant) as the best estimator among the family of estimators $\{\hat{\Psi}_\alpha : \alpha \in \mathcal{A}\}$ when applied to samples of size $n(1 - 1/V)$, or achieve the parametric model rate $1/n$ up till a $\log n$ factor. By a simple argument as presented in van der Laan and Dudoit (2003), Dudoit and van der Laan (2005) and van der Vaart et al. (2006), it follows that by letting the $V = V_n$ in the V-fold cross-validation scheme converge to infinity at a slow enough rate relative to n, then either $\psi_n = \hat{\Psi}_{\alpha_n}(P_n)$ performs asymptotically as well (up till the constant) as the best estimator among the estimators $\{\hat{\Psi}_\alpha : \alpha\}$ applied to the full sample P_n, or it achieves the parametric rate of convergence up till the $\log n$ factor.

The take home message of this theorem is that our super learner will perform asymptotically as well as the best learner among the family of candidate learners $\hat{\Psi}_\alpha$ indexed by α. By choosing the regression model $m(\cdot \mid \alpha)$ so that there exist a α_j so that $m(Z \mid \alpha_j) = Z_j$ for each $j = 1, \ldots, J$ (e.g., $m(Z \mid \alpha) = \alpha Z$), then it follows, in particular, that the resulting prediction algorithm asymptotically outperforms each of the initial candidate estimators $\hat{\Psi}_j$. More importantly and practically, the set of candidate estimators $\hat{\Psi}_\alpha$ can include interesting combinations of these J estimators which exploit the strengths of various of these estimators for the particular data generating distribution P_0 instead of focusing on one of them. For example, if one uses the linear regression model $m(Z \mid \alpha) = \alpha Z$, then the candidate estimators $\{\hat{\Psi}_\alpha : \alpha\}$ include all averages of the J estimators, including convex combinations. As becomes evident in our data analysis and simulation results, the selected super learner ψ_n^* based on a linear (or logistic) regression model is often indeed (or logistic function of) a weighted average of competing estimators in which various of the candidate learners significantly contribute to

the average.

4 Simulation results

In this section, we conducted 3 simulation studies to evaluate the working characteristics of the super learner. These simulations all involve a continuous response variable. For the first simulation, the true model is:

$$\begin{aligned} Y_i = \; & 2w_1w_{10} + 4w_2w_7 + 3w_4w_5 - 5w_6w_{10} + 3w_8w_9 + w_1w_2w_4 \\ & -2w_7(1-w_6)w_2w_9 - 4(1-w_{10})w_1(1-w_4) + \varepsilon \end{aligned} \tag{3}$$

where $w_j \sim Binomial(p = 0.4)$, $j = 1, \ldots, 10$ and $\varepsilon \sim Normal(0, 1)$. Each observation consists of the 10 dimensional covariate vector W, and the continuous response variable Y. The parameter of interest is $\psi_0(W) = \mathrm{E}_0(\mathrm{Y}|\mathrm{W})$. The simulated learning data set contains a sample of 500 observations (i=1,...,500) from model 3.

We applied the super learner to the learning set using five candidate learners. The first candidate was a simple linear regression model with only main terms, which will be estimated with regular least squares. The second candidate was main terms LARS. Internal cross-validation (i.e. another layer of cross-validation inside each training split) was used the estimate the optimal fraction parameter, $\lambda_0 \in (0, 1)$. The third candidate was the D/S/A algorithm for data-adaptive polynomial regression. For the D/S/A algorithm, we allowed interaction terms and restricted the model to less than 50 terms. The D/S/A uses internal cross-validation to determine the best model in this model space. The fourth candidate was logic regression where the number of trees was selected to be 5 and the number of leaves to be 20 based on 10-fold cross validation of the learning data set. For the logic regression fine-tuning parameters, we searched over #trees $\in \{1, \ldots, 5\}$ and #leaves $\in \{1, \ldots, 20\}$. The final candidate algorithm was random forests. Table 1 contains references for the R packages of each canidate learner.

We applied the super learner with 10-fold cross-validation on the learning set. Applying the prediction to all 10 folds of the learning set gives us the predicted values $Z_i \equiv (\hat{\Psi}_{j\nu(i)}(W_i) : j = 1, \ldots, 5)$ and corresponding Y_i for each observation $i = 1, \ldots, 500$. We then proposed the linear model $\mathrm{E}(\mathrm{Y}|\mathrm{Z}) = \alpha + \beta\mathrm{Z}$ and used least squares to estimate the intercept α and parameter vector β based on (Y_i, Z_i), $i = 1, \ldots, n$.

method	RMSPE	β_n
Least Squares	1.00	0.038
LARS	1.15	-0.171
D/S/A	0.22	0.535
Logic	0.32	0.274
Random Forest	0.42	0.398
Super Learner	0.20	

Table 2: Simulation Example 1: Estimates of the relative mean squared prediction error (compared to least squares) based on a learning sample of 500 observations and the evaluation sample M=10,000. The estimates for β in the super learner are also reported in the right column ($\alpha_n = -0.018$).

After having obtained the fit α_n, β_n of α, β, next, each of the candidate learners was fit on the entire learning set to obtain $Psi_j(P_n)(W)$, which gives the super learner $\hat{\Psi}(P_n)(W) = \alpha_n + \beta_n(\hat{\Psi}_j(P_n)(W) : j = 1, \ldots, 5))$ when applied to a new covariate vector W.

To evaluate the super learner next to each of the candidate learners, an additional 10,000 observations are simulated from the same data generating distribution. This new sample is denoted the evaluation sample. Using the models on the learning data set, we calculated the mean squared prediction error (MSPE) on this new evaluation data set for the super learner and each of the candidate learners. Table 2 has the results for the relative mean squared prediction error (RMPSE), where $RMSPE(x) = MSPE(x)/MSPE(least\ squares)$. Among the candidate learners, the D/S/A algorithm appears to have the smallest error, but the super learner improves on the D/S/A fit. The estimates β_n all appear to be nonzero except for the simple linear regression model. The super learner can combine information from the candidate learners to build a better predictor.

The second simulation considers continuous covariates as opposed to binary covariates from the first simulation. Let X be a 20 dimensional multivariate normal random vector and $X \sim N_p(0, 16 * Id_p)$ where $p = 20$ and Id_p is the p-dimensional identity matrix. Each column of X is a covariate in the models used below. The outcome is defined as:

$$Y_i = X_1X_2 + X_{10}^2 - X_3X_{17} - X_{15}X_4 + X_9X_5 + X_{19} - X_{20}^2 + X_9X_8 + \varepsilon, \quad (4)$$

where $\varepsilon \sim Normal(0, 16)$ and X_j is the j^{th} column of X. From this model, 200 observations were simulated for the learning data set and an additional

5,000 were simulated for the evaluation data set similar to the first simulation. The super learner was applied with the following candidate learners:

- Simple linear regression with all 20 main terms.
- LARS with internal cross-validation to find the optimal fraction.
- D/S/A with internal cross-validation to select the best model with fewer than 25 terms allowing for interaction and quadratic terms.
- Ridge regression with internal cross-validation to select the optimal $L2$ penalty parameter.
- Random forests with 1,000 trees.
- Adaptive regression splines.

Table 3 contains the results for the second simulation. As in the first simulation, the relative mean squared prediction error is used to evaluate the candidate learners and the super learner. For this model, simple linear regression, LARS, and ridge regression all appear to have the same results. Random forests and adaptive regression splines are better able to pick up the non-linear relationship, but among the candidate learners, the D/S/A is the best with a relative MSPE of 0.43. But the super learner improves on the fit even more with a relative MSPE of 0.22 by combining the candidate learners. Since the model for $\psi_n^*(z)$ can be near collinear, the estimates of β are often unstable and should not be used to determine the best candidate by comparing the magnitude of the parameter estimate.

The main advantage of the proposed super learner is the adaptivity to different data generating distributions across many studies. The third simulation demonstrates this feature by creating 3 additional studies and applying the super learner and the candidates to all 3 studies then combining the results with the second simulation and evaluating the mean square error across all 4 studies. Equation 5 shows the data generating distributions for the 3 new studies. The data generating distribution for the covariates X is the same as the second simulation example above. To be consistent across the 4 studies, the same candidate learners from the second simulation were applied to these 3 new studies.

method	RMSPE	β_n
Least Squares	1.00	-0.73
LARS	0.91	-0.92
D/S/A	0.43	0.86
Ridge	0.98	0.61
Random Forest	0.71	1.06
MARS	0.61	0.05
Super Learner	0.22	

Table 3: Simulation Example 2: Estimates of the relative mean squared prediction error (compared to Least Squares) based on a learning sample of 200 observations and the evaluation sample M=5,000. The estimates for β in the super learner are also reported in the right column ($\alpha_n = 0.03$).

$$Y_{ij} = \begin{cases} -5 + X_2 + 6(X_{10}+8)_+ - 6(X_{10})_+ - 7(X_{10}-5)_+ \\ \quad - 6(X_{15}+6)_+ + 8(X_{15})_+ + 7(X_{15}-6)_+ + \varepsilon & \text{if } j = 1 \\ 10 \cdot \mathrm{I}(X_1 > -4 \; and \; X_2 > 0 \; and \; X_3 > -4) + \varepsilon & \text{if } j = 2 \\ -4 + X_2 + \sqrt{|X_3|} + sin(X_4) - .3X_6X_{11} + 3X_7 \\ \quad + .3X_8^3 - 2X_9 - 2X_{10} - 2X_{11} + \varepsilon & \text{if } j = 3 \end{cases} \quad (5)$$

where $\varepsilon \sim Normal(0, 16)$ and $\mathrm{I}(x) = 1$ if x is true, and 0 otherwise. For the 4 studies (the 3 new studies combined with the second simulation), the learning sample contained 200 observations and the evaluation sample contained 5,000 observations.

Table 4 contains the results from the second simulation. For the first study ($j = 1$), the adaptive regression spline function is able to estimate well the true distribution. The super learner is not able to improve on the fit, but it does not do worse than the best candidate algorithm. In the second study ($j = 2$), the adaptive regression spline function is not the best candidate learner. The random forests performs best in the second study, but the super learner is able to improve on the fit. The third study ($j = 3$) is similar to the first in that the adaptive regression splines function is able to approximate the true distribution well, but the super learner does not do worse. The squared prediction error from these three studies and the second

method	study 1	study 2	study 3	2^{nd} simulation	overall
Least Squares	1.00	1.00	1.00	1.00	1.00
LARS	0.91	0.95	1.00	0.91	0.95
D/S/A	0.22	0.95	1.04	0.43	0.71
Ridge	0.96	0.99	1.02	0.98	1.00
Random Forest	0.39	0.72	1.18	0.71	0.91
MARS	0.02	0.82	0.17	0.61	0.38
Super Learner	0.02	0.67	0.16	0.22	0.19

Table 4: Simulation Example 3: Estimates of the relative mean squared prediction error (compared to least squares) based on the validation sample. The 3 new studies from 5 are combined with the second simulation example and the relative mean squared prediction error is reported in the overall column.

simulation was combined to give a mean squared prediction error for the four studies. The last column in table 4 gives the relative mspe for each of the candidate learners and the super learner. If the researcher had selected just one of the candidate learners, they might have done well within one or two of the studies, but overall the super learner will outperform the candidate learners. For example, the MARS learner performs well on the first and third study, and does well overall with a relative MSPE of 0.38, but the super learner outperforms the MARS learner with an overall relative MSPE of 0.19. The super learner is able to adapt to the different data generating distributions and will outperform any candidate learner across many studies.

5 Data Analysis

We applied the super learner to the diabetes data set from the LARS package in R. Details on the data set can be found in Efron et al. (2004). The data set consists of 442 observations of 10 covariates (9 quantitative and 1 qualitative) and a continuous outcome. The covariates have been standardized to have mean zero and unit L2 norm. We selected 6 candidate learners for the super learner. The first candidate was least squares using all 10 covariates. Next we considered the least squares model with all possible two-way interactions and quadratic terms on the quantitative covariates. The third and fourth candidates were applying LARS to the main effects and all possible two-

way interaction models above. Internal cross-validation was used to select the "fraction" point for the prediction. The fifth candidate algorithm was D/S/A allowing for two-way interactions and a maximum model size of 64. The final candidate learner was the random forests algorithm. For the super learner, we then used a linear model and estimated the parameters with least squares.

We also applied the proposed super learner to the HIV-1 drug resistance data set in Sinisi et al. (2007) and Rhee et al. (2006). The goal of the data is to predict drug susceptibility based on mutations in the protease and reverse transcriptase enzymes. The HIV-1 sequences were obtained from publicly available isolates in the Stanford HIV Reverse Transcriptase and Protease Sequence Database. Details on the data and previous analysis can be found in Sinisi et al. (2007) and Rhee et al. (2006). The outcome of interest is standardized log fold change in drug susceptibility, defined as the ratio IC_{50} of an isolate to a standard wildtype control isolate; IC_{50} (inhibitory concentration) is the concentration of the drug needed to inhibit viral replication by 50%. We focused our analysis to a single protease inhibitor, nelfinavir, where we have 740 viral isolates in the learning sample of 61 binary predictor covariates and one quantitative outcome.

For the HIV data set, we considered six candidate learners. The first candidate was least squares on all main terms. The second candidate was the LARS algorithm. Internal cross validation was used to determine the best fraction parameter. The third candidate was logic regression. Similar to the simulation example, we used 10-fold cross-validation on the entire learning set to determine the parameters,#trees $\in \{1, \ldots, 5\}$ and #leaves $\in \{1, \ldots, 20\}$, for logic regression. For the HIV data set, we selected #trees = 5 and #leaves = 10. The fourth candidate was the CART algorithm. We also applied the D/S/A algorithm searching over only main effects terms and a maximum model size of 35. The final candidate was random forests. For the super learner, a linear model was used to estimate the parameters with least squares. All models were fit in R similar to the simulation example above.

To evaluate the performance of the super learner in comparison to each of the candidate learners we split the learning data set into 10 validation data sets and corresponding training data sets. The super learner and each candidate learner was fit one each fold of the cross-validation, giving us a honest cross-validated risk estimate to compare the super learner to each of the candidate learners.

Method	RCV risk	β_n
Least Squares (1)	1.00	0.172
Least Squares (2)	1.13	-0.003
LARS (1)	1.07	0.239
LARS (2)	1.08	0.126
D/S/A	0.98	0.481
Random Forests	1.07	0.027
Super Learner	0.98	

Table 5: Super learner results for the diabetes data set. Least Squares (1) and LARS (1) refer to the main effects only models. Least Squares (2) and LARS (2) refer to the all possible two-way interaction models. Relative 10-fold Honest Cross-Validation risk estimates, compared to main terms least squares (RCV risk) are reported. β_n in the super learner is reported in the last column ($\alpha_n = -6.228$).

5.1 Super Learner Results

Table 5 presents results for the diabetes data analysis. A 10-fold cross-validation estimate of the mean squared error was calculated, and the relative risk estimate is reported. The relative cross-validation risk estimate (RCV) is $RCV(x) = CV(x)/CV(\textit{main terms least squares})$, where $CV(x)$ is the cross-validation risk estimate for x. Based on the cross validated estimate, the D/S/A has the best estimate among the candidate learners. The super learner does not appear to improve significantly on the D/S/A learner, but it does not do any worse either. We also report the estimates α_n and β_n used in the super learner. The D/S/A algorithm has the largest coefficient (0.481) and appears to be given the most weight in the super learner. We also note that least squares with all possible two-way interactions is barely used in the super learner, with a coefficient of -0.003. This example shows how the super learner can use cross validation to data adaptively select (i.e. give more weight) to the better candidate predictors.

Table 6 presents the results for the HIV data analysis. Based on 10-fold cross validated estimates of the mean squared error, main terms least squares performs best, although random forests and LARS have similar error estimates to least squares. In contrast to the diabetes data analysis above, D/S/A does not perform well on this data set. This highlights the need for a super learner since one candidate algorithm will not work on all data

Method	RCV risk	β_n
Least Squares	1.00	0.552
LARS	1.03	0.075
Logic	1.52	-0.020
CART	1.77	0.076
D/S/A	1.53	-0.161
Random Forests	1.02	0.510
Super Learner	0.87	

Table 6: Super learner results for the HIV data set. Relative 10-fold honest cross validated risk estimates (RCV risk) compared to least squares are reported. β_n in the super learner is reported in the last column ($\alpha_n = 0.027$).

sets. Among the candidate learners, least squares has the smallest cross-validated risk estimate, but the super learner has a smaller risk estimate ($RCV = 0.87$). We also present the estimates for α and β in table 6. Both least squares and random forests appear to be receiving the most weight in the super learner with coefficients 0.552 and 0.510 respectively. Again, the super learner can use the cross validated predictions to data adaptively build the best predictor.

These are both situations where one of the candidate learners does a good job of prediction and gives little room for improvement for the super learner. But these examples also demonstrate that one candidate algorithm may not be flexible enough to perform best on all data generating distributions and since a researcher is unlikely to know *a priori* which candidate learner will work best, the super learner is a natural choice for prediction.

6 Discussion.

The new super learning approach provides both a fundamental theoretical as well as practical improvement to the construction of a predictor. The super learner is a flexible prediction algorithm which can perform well on many different data generating distributions, and utilizes cross-validation to protect against over-fitting. We wish to stress that the theory suggests that to achieve the best performance one should not apply this algorithm to a restricted set of candidate learners, but one should aim to include any available sensible learners. In addition, the amount of computations does

not exceed the amount of computations it takes to calculate each of the candidate learners on the training and full data sets. In our simulations we used a particular set of available learners only because they were easily available as R functions. Thus, the potential for improving learners applies to a very wide array of practical problems.

Our results generalize to parameters which can be defined as minimizers of a loss function, including (unknown) loss functions indexed by parameters of the true data generating distribution (van der Laan and Dudoit (2003)). In particular, the super learner approach applies to maximum likelihood estimation in semiparametric or nonparametric models for the data generating distribution, and to targeted maximum likelihood estimation with respect to a particular smooth functional of the density of the data, as presented in van der Laan and Rubin (2007).

7 Appendix

Under the Assumption A1 that the loss function $L(O,\psi) = (Y - \psi(X))^2$ is uniformly bounded, and the Assumption A2 that the variance of the ψ_0-centered loss function $L(O,\psi) - L(O,\psi_0)$ can be bounded by its expectation uniformly in ψ, van der Laan et al. (2006) (Theorem 3.1) establish the following finite sample inequality.

Theorem 2 *Let* $\{\hat{\psi}_k = \hat{\Psi}_k(P_n), k = 1, ..., K(n)\}$ *be a given set of* $K(n)$ *estimators of the parameter value* $\psi_0 = \arg\min_{\psi\in\Psi} \int L(o,\psi)dP_0(o)$. *Let* $d_0(\psi,\psi_0) \equiv E_{P_0}\{L(O,\psi) - L(O,\psi_0)\}$ *denote the risk difference between a candidate estimator* ψ *and the parameter* ψ_0. *Suppose that* $\mathbf{\Psi}$ *is a parameter space so that* $\hat{\Psi}_k(P_n) \in \mathbf{\Psi}$ *for all* k, *with probability 1. Let* $\hat{K}(P_n) \equiv \arg\min_k E_{B_n} \int L(o, \hat{\Psi}_k(P^0_{n,B_n}))dP^1_{n,B_n}(o)$ *be the* cross-validation selector, *and let* $\tilde{K}(P_n) \equiv \arg\min_k E_{B_n} \int L(o, \hat{\Psi}_k(P^0_{n,B_n}))dP_0(o)$ *be the comparable* oracle selector. *Let* p *be the proportion of observations in the validation sample. Then, under assumptions A1 and A2, one has the following finite sample inequality for any* $\lambda > 0$ *(where* $C(\lambda)$ *is a constant, defined in van der Laan et al. (2006)):*

$$Ed_0(\hat{\Psi}_{\hat{K}(P_n)}(P^0_{n,B_n}),\psi_0) \leq (1+2\lambda)Ed_0(\hat{\Psi}_{\tilde{K}(P_n)}(P^0_{n,B_n}),\psi_0)+2C(\lambda)\frac{1+log(K(n))}{np}$$

References

L. Breiman. Random Forests. *Machine Learning*, 45:5–32, 2001.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. The Wadsworth Statistics/Probability series. Wadsworth International Group, 1984.

S. Dudoit and M. J. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2:131–154, 2005.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least Angle Regression. *Annals of Statistics*, 32(2):407–499, 2004.

Y. Freund, R. E. Schapire, Y. Singer, and M. K. Warmuth. Using and combining predictors that specialize. In *Twenty-Ninth Annual ACM Symposium on the Theory of Computing*, 1997.

J. H. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19(1):1–141, 1991.

J. V. Hansen. Combining predictors: Some old methods and a new method. In *ICCIN*, 1998. URL `citeseer.ist.psu.edu/article/hansen98combining.html`.

A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(3):55–67, 1970.

S. Rhee, J. Taylor, G. Wadhera, J. Ravela, A. Ben-Hur, D. Brutlag, and R. W. Shafer. Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences USA*, 2006.

I. Ruczinski, C. Kooperberg, and M. LeBlanc. Logic Regression. *Journal of Computational and Graphical Statistics*, 12(3):475–511, 2003.

S. E. Sinisi and M. J. van der Laan. Deletion/Substitution/Addition algorithm in learning with applications in genomics. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004. Article 18.

S. E. Sinisi, E. C. Polley, S.Y. Rhee, and M. J. van der Laan. Super learning: An application to the prediction of HIV-1 drug resistance. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007.

M. J. van der Laan and S. Dudoit. Unified Cross-Validation Methodology for Selection Among Estimators and a General Cross-Validated Adaptive Epsilon-Net Estimator: Finite Sample Oracle Inequalities and Examples. Technical Report 130, Division of Biostatistics, University of California, Berkeley, Nov. 2003. URL `http://www.bepress.com/ucbbiostat/paper130/`.

M. J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *International Journal of Biostatistics*, 2(1), 2007.

M. J. van der Laan, S. Dudoit, and A. W. van der Vaart. The cross-validated adaptive epsilon-net estimator. *Statistics and Decisions*, 24(3):373–395, 2006.

A.W. van der Vaart, S. Dudoit, and M.J. van der Laan. Oracle inequalities for multi-fold cross vaidation. *Statistics and Decisions*, 24(3), 2006.

3.2 *Loss-Based Cross-Validated Deletion/Substitution/Addition Algorithms in Estimation*

The following article appears as it was originally published on the University of California, Berkeley Division of Biostatistics Working Paper Series website in 2004, http://www.bepress.com/ucbbiostat/paper143/.

It was later published as *Deletion/Substitution/Addition Algorithm in Learning with Applications in Genomics* in *Statistical Applications in Genetics and Molecular Biology* in 2004, http://www.bepress.com/sagmb/vol3/iss1/art18/.

Loss-Based Cross-Validated Deletion/Substitution/Addition Algorithms in Estimation

Sandra E. Sinisi, Mark J. van der Laan
Division of Biostatistics, University of California, Berkeley

Abstract

In van der Laan and Dudoit (2003) we propose and theoretically study a unified loss function based statistical methodology, which provides a road map for estimation and performance assessment. Given a parameter of interest which can be described as the minimizer of the population mean of a loss function, the road map involves as important ingredients cross-validation for estimator selection and minimizing over subsets of basis functions the empirical risk of the subset-specific estimator of the parameter of interest, where the basis functions correspond to a parameterization of a specified subspace of the complete parameter space. In this article we first review this approach. Then we propose a general deletion/substitution/addition algorithm for minimizing over subsets of variables (e.g., basis functions) the empirical risk of subset-specific estimators of the parameter of interest. In particular, in the regression context, this algorithm corresponds to minimizing over subsets of variables the sum of squared residuals of the subset-specific linear regression estimator. This algorithm provides us with a new class of loss-based cross-validated algorithms in prediction of univariate and multivariate outcomes, conditional density and hazard estimation, and we generalize it to censored outcomes such as survival. In the context of regression, using polynomial basis functions, we study the properties of the deletion/substitution/addition algorithm in simulations and apply the method to detect binding sites in yeast gene expression experiments.

1 Introduction.

This article introduces the Deletion/Substitution/Addition (D/S/A) algorithm by demonstrating how it works in linear regression with polynomial basis functions and illustrates the utility of this method in genomics by applying it to the detection of binding sites in a publicly available dataset of the yeast *Saccharomyces cerevisiae*. The D/S/A algorithm can be used in a variety of settings including prediction of univariate outcomes (setting used in this article), prediction of multivariate outcomes, conditional density and hazard estimation, and can be generalized to censored outcomes such as survival. Our motivation stems from current statistical inference problems in the analysis of genomic data, such as the prediction of biological and clinical outcomes using microarray gene expression measures, the identification of regulatory motifs (i.e., transcription factor binding sites) in DNA sequences, and the genetic mapping of complex traits using single nucleotide polymorphisms (SNPs). One such motivating example is the identification of regulatory motifs in DNA sequences. Transcription factors (TF) are proteins that selectively bind to DNA to regulate gene expression. The transcription factor binding sites, or regulatory motifs, are short DNA sequences (5-25 base pairs) in the upstream control region (UCR) of genes, i.e., in regions roughly 600 to 1,000 base pairs from the gene start site (in lower eukaryotes, e.g., yeast). A possible statistical question is to utilize gene expression data to identify sequence motifs associated with genes that are activated under a specified experimental condition. Thus, the estimation problem can be framed as a prediction problem where one is predicting gene expression levels based on sequence features, such as pentamers (and their interactions). Making statistical inferences based on voluminous, genomic data involves exploring many different relationships (to account for high-order interactions) amongst a vast array of explanatory variables and a single outcome or multiple outcomes. As a result, dominating features of statistical inference problems in genomics include a high-dimensional parameter space and deciding upon a reasonable error measure which we wish to minimize. When faced with a high-dimensional parameter space, it becomes difficult to minimize a suitable error measure over the entire parameter space. A unified framework to approach these problems has been offered by van der Laan and Dudoit (2003).

Given a large parameter space, we want to perform intensive searches over this space and form candidate estimators which address the desired

statistical question. Once we collect all these candidate estimators, we must then select a *best* estimate. The approach of van der Laan and Dudoit (2003) establishes that cross-validation can be used to select among many candidate estimators, even in finite sample situations. Furthermore, it is shown that aggressive searches are adaptive to the truth. van der Laan et al. (2004) propose a cross-validated adaptive ϵ-net estimation methodology where they consider collections of subspaces of the parameter space. For each choice of subspace and resolution, they generate candidate estimators as the empirical risk minimizers over ϵ-*nets*.

Consequently, it becomes necessary to construct an algorithm that is capable of adapting to the data completely to address inference questions satisfactorily. This means that we need an algorithm that is capable of minimizing a suitable error measure, say the empirical mean of a loss function, over an arbitrarily good approximation of the complete parameter space, and we need cross-validation to select certain fine-tuning parameters. However, we do not want a nonparametric version of such an algorithm because it will try to fit the data perfectly resulting in estimators that are too variable due to the number of variables involved relative to a limited sample size. Consequently, we want to put certain stops on the algorithm and index it by a number of "brakes." These brakes correspond to specified subspaces of the complete parameter space. A natural collection of brakes can be obtained by parameterizing the complete parameter space in terms of linear combinations of basis functions where the choice of a basis, the number of basis functions, complexity measure(s) on the basis functions, and a constraint on the vector of coefficients (e.g., norm) provide natural choices for brakes. Based on the cross-validation results given by van der Laan and Dudoit (2003), even when one implements a large number of brakes, the resulting estimator will perform asymptotically exactly as well as the estimator corresponding to the oracle selector of brakes. As a consequence, the estimator adapts at an asymptotically increasing level to the truth when more and more brakes are applied. Hence, such notions as aggressive searches, using basis functions to parameterize large parameter spaces, applying certain brakes, and cross-validation selection led us to construct algorithms which incorporate these ideas to answer statistical questions in genomics.

In this article, we propose a D/S/A algorithm for minimizing empirical risk over a subspace. In the context of regression, we will study the performance of the D/S/A algorithm in simulation studies. Finally, we apply the methodology to a yeast data set to detect binding sites.

2 Review of Estimation Road Map.

This section briefly explains the concepts of loss-function based estimation and cross-validation selection which are important components of the D/S/A algorithm. Details on theoretical aspects behind the D/S/A algorithm are available in Sinisi and van der Laan (2004) and Dudoit et al. (2003). A more thorough description of loss-function based estimation is available in van der Laan and Dudoit (2003) and van der Laan et al. (2004).

Let $(O, \psi) \rightarrow L(O, \psi) \in \mathbb{R}$ be a (loss) function which maps a candidate parameter value $\psi \in \mathbf{\Psi}$ and observation O into a real number. The expectation of this loss function is minimized at ψ_0:

$$\begin{aligned} \psi_0 &= \text{argmin}_{\psi \in \mathbf{\Psi}} \int L(o, \psi) dP_0(o) \\ &= \text{argmin}_{\psi \in \mathbf{\Psi}} E_0 L(O, \psi). \end{aligned} \tag{1}$$

In univariate outcome regression, we have $O = (Y, W) \sim P_0$, where Y is a scalar outcome and W is a vector of covariates. The parameter of interest is the conditional expected value, $\psi_0(W) \equiv E_{P_0}(Y \mid W)$, of the outcome Y given covariates W. We can use as loss function the *quadratic loss function*:

$$L(O, \psi) \equiv L(Y, W, \psi) = (Y - \psi(W))^2,$$

also known as the *squared error loss function* or the L^2 *loss function*.

Parameterization of the parameter space.

Having defined the parameter of interest as the risk minimizer for the squared error loss function, the next task is to generate a sequence of candidate estimators by minimizing the empirical risk over subspaces of increasing dimension approximating the complete parameter space $\mathbf{\Psi}$. We propose to parameterize $\mathbf{\Psi}$ in terms of tensor products of basis functions with polynomial basis functions. Given a d-vector $\vec{p} = (p_1, \ldots, p_d) \in \mathbb{N}^d$, we denote the polynomial basis functions by $\phi_{\vec{p}}(W) = W_1^{p_1} \ldots W_d^{p_d}$ where the collection $\{\phi_{\vec{p}} : \vec{p} \in \mathbb{N}^d\}$ provides a basis for the complete parameter space $\mathbf{\Psi}$. The index set $I \subset \mathcal{I}$ represents a set of elements in $\mathbb{N}^d$.

Next, we define a collection of subspaces $\mathbf{\Psi}_s \subset \mathbf{\Psi}$, indexed by s ranging over a set $\mathcal{A}_n$. Such subspaces can be obtained by restricting the subsets I

of basis functions to be contained in $\mathcal{I}_s \subset \mathcal{I}$:

$$\mathbf{\Psi}_s = \left\{ \sum_{\vec{p} \in I} \beta_{\vec{p}} \phi_{\vec{p}} \in \mathbf{\Psi} : m(I) \leq s \right\}, \tag{2}$$

where $m(I) = (m_1(I), \ldots, m_q(I))$ is defined as a q-valued function such that $m_1(I) \leq s_1, \ldots, m_q(I) \leq s_q$. In this article, $m_1(I) = |I|$ represents the number of tensor products or the size of the index sets, $m_2(I) = \max_{\vec{p} \in I} \sum_{j=1}^{d} I(p_j \neq 0)$ represents the maximum order of interaction of tensor products (the number of non-zero components in $\vec{p}$), and $m_3(I) = \max_{\vec{p} \in I} \sum_{j=1}^{d} p_j$ represents the maximum sum of powers of tensor products.

Construction of candidate estimators.

After having defined our subspaces, $\mathbf{\Psi}_s$, we would like to find the minimizer of the empirical risk over the subspace for each $s \in \mathcal{A}_n$. This minimization problem is naturally split into two sequential steps. Given each possible subset $I \in \mathcal{I}_s$ of basis functions, compute the corresponding minimum risk estimator of β, which in regression corresponds to minimizing the sum of the squared residuals over the linear regression model in the basis functions $\phi_{\vec{p}}$ indexed by $\vec{p} \in I$. For each I this results in an estimator $\Psi_{I,s}(P_n) \equiv \psi_{I,\beta(P_n|I,s)}$.

Now, it remains to minimize the empirical risk over all allowed subsets $I \in \mathcal{I}_s$ of basis functions. Specifically, one needs to minimize the function $f_{E,s} : \mathcal{I}_s \to \mathbb{R}$ defined by

$$f_{E,s}(I) \equiv \int L\left(O, \Psi_{I,s}(P_n)\right) dP_n(O). \tag{3}$$

Let

$$I_s(P_n) \equiv \text{argmin}_{I \in \mathcal{I}_s} f_{E,s}(I)$$

be the minimizer. In Section 3 we propose a D/S/A algorithm that seeks to calculate $I_s(P_n)$.

Selection among candidate estimators: Cross-validation.

Now, we have the empirical risk minimizer, denoted by $\hat{\Psi}_s(P_n)$, for each choice of subspace s. The final task is to select s with cross-validation.

To derive a general representation for cross-validation, let $B_n \in \{0,1\}^n$ be a random vector whose observed value defines a split of the observed data $O_1, \ldots, O_n$, the learning sample, into a validation sample and a training sample. If $B_n(i) = 0$ then observation i is placed in the training sample and if $B_n(i) = 1$, it is placed in the validation sample. We will denote the empirical distribution of the data in the training sample and validation sample with P^0_{n,B_n} and P^1_{n,B_n}, respectively. The proportion of observations in the validation sample is denoted by $p = \sum_i B_n(i)/n$. The cross-validation selector of s is now defined as

$$\begin{aligned} s(P_n) &\equiv \text{argmin}_{s \in \mathcal{A}_n} E_{B_n} \int L(O, \hat{\Psi}_s(P^0_{n,B_n})) dP^1_{n,B_n}(O) \\ &= \text{argmin}_{s \in \mathcal{A}_n} E_{B_n} \frac{1}{np} \sum_{i=1}^{n} I(B_n(i) = 1) L(O_i, \hat{\Psi}_s(P^0_{n,B_n})). \end{aligned}$$

Our final estimator of our parameter of interest is given by $\hat{\Psi}(P_n) \equiv \Psi_{s(P_n)}(P_n)$.

For the finite sample inequalities comparing the risk distance of the cross-validation selected estimator with the risk distance of the estimator chosen by the oracle selector and its asymptotic implications, we refer to van der Laan and Dudoit (2003) for results for general loss functions, and Dudoit and van der Laan (2003), van der Laan et al. (2003) for the corresponding results in regression and likelihood cross-validation. The practical message of these results is that for quadratic (e.g., convex) uniformly bounded loss functions the cross-validation selector performs as well in risk distance as the oracle selector up to a term smaller than $C \log(K(n))/(np)$, while for non-quadratic loss functions, this last term is replaced by $C\sqrt{\log(K(n)/(np)}$. Thus, as long as the number $K(n)$ of estimators we consider is such that $\log(K(n))/(np)$ is of smaller order than the actual minimal risk distance $\min_{s \in \mathcal{A}_n} d(\hat{\psi}_s, \psi_0)$ of the candidate estimators to ψ_0, then the cross-validation selector is asymptotically equivalent (in risk distance) to the oracle selector. That is, in estimation problems which do not allow the parametric $1/\sqrt{n}$-rate of convergence, the number $K(n)$ of candidate estimators can be a polynomial power in n.

3 D/S/A algorithm for minimizing over subsets of basis functions.

In this section, we propose an aggressive and flexible algorithm for generating a sequence of index sets I, according to three types of moves for the elements of I: deletions, substitutions, and additions. We refer to this general algorithm as the *Deletion/Substitution/Addition algorithm*, or *D/S/A algorithm*. The main features of this approach are summarized below for the case where the index sets are subsets of $\mathbb{N}^d$, as is the case for tensor product polynomial basis functions $\phi_{\vec{p}}$. The D/S/A algorithm has been adapted to histogram regression with partition-specific indicator basis functions as provided by Molinaro and van der Laan (2004).

To simplify notation, in this section we will suppress dependence of quantities on s; let $\mathcal{I}$ denote the collection of allowed index sets. Let s_0 denote the dimension d and assume that $s = (s_1, \ldots, s_q)$, where s_1 denotes the upper bound on the size of the index sets (i.e, the number of allowed basis functions), while $s_2, \ldots, s_q$ represent the remaining fine tuning parameters. The D/S/A algorithm described below aims to calculate $I_s(P_n)$ for each choice of s_1, given the remaining components of s. Thus, one has to carry out this algorithm for each value of $s_2, \ldots, s_q$ to obtain all optimal index sets $\{I_s(P_n) : s\}$ and thereby our collection of s-specific estimators $\hat{\Psi}_s(P_n)$. Throughout this section, we use k to represent s_1 where $s_1 = |I|$.

The D/S/A algorithm for minimizing over index sets I is defined in terms of three functions, $DEL(I)$, $SUB(I)$, and $ADD(I)$, which map an index set $I \in \mathcal{I}$ of size k into *sets of index sets* of size $k-1$, k, and $k+1$, respectively.

Deletion/Substitution/Addition moves.
Consider index sets $I \subset \mathbb{N}^d$ and let $\mathcal{I}$ denote a collection of subsets of $\mathbb{N}^d$.
Deletion moves. Given an index set $I \in \mathcal{I}$ of size $k = |I|$, define a set $DEL(I) \subset \mathcal{I}$ of index sets of size $k-1$, by deleting individual elements of I. This results in k possible deletion moves, i.e., $|DEL(I)| = k$.
Substitution moves. Given an index set $I \in \mathcal{I}$ of size $k = |I|$, define a set $SUB(I) \subset \mathcal{I}$ of index sets of size k, by replacing individual elements $\vec{p} \in I$ by one of the $2d$ vectors created by adding or subtracting 1 to any of the d components of $\vec{p}$. That is, for each $\vec{p} \in I$, consider moves $\vec{p} \pm \vec{u}_j$, where $\vec{u}_j$ denotes the unit d-vector with one in position j and zero elsewhere, $j = 1, \ldots, d$. This results in up to $k \times (2d)$ possible substitution moves, i.e., $|SUB(I)| = k \times (2d)$. In case the allowed index sets require that each $\vec{p}$ has at most s_2 non-zero components, then we propose to add to these sub-

stitution moves the *alternate-substitution* moves. The alternate-substitution moves correspond to adding or subtracting the unit vectors as above, but if that results in a $\vec{p}$ with more than s_2 non-zero components, then we replace it by the s_2 vectors one obtains by setting one of the (original) non-zero components equal to zero. This augmentation of the set of substitution moves results in maximally $k \times s_2 \times (2d)$ substitution moves.

Addition moves. Given an index set $I \in \mathcal{I}$ of size $k = |I|$, define a set $ADD(I) \subset \mathcal{I}$ of index sets of size $k+1$, by adding to I an element of $SUB(I)$ or one of the d unit vectors $\vec{u}_j$, $j = 1, \ldots, d$. This results in up to $k \times (2d) + d$ (or $k \times s_2 \times (2d) + d$) possible addition moves, i.e., $|ADD(I)| = k \times (2d) + d$.

Thus the substitution moves (excluding the alternate-substitution moves) can be described as

$$SUB(I) \to \begin{cases} (p_1 + 1, p_2, p_3, \ldots, p_d) \\ (p_1, p_2 + 1, p_3, \ldots, p_d) \\ \vdots \\ (p_1, p_2, p_3, \ldots, p_d + 1) \\ (p_1 - 1, p_2, p_3, \ldots, p_d) \\ (p_1, p_2 - 1, p_3, \ldots, p_d) \\ \vdots \\ (p_1, p_2, p_3, \ldots, p_d - 1) \end{cases}$$

for each $\vec{p} \in I$, and the addition moves as adding $\vec{p}_{k+1}$ described by

$$ADD(I) = \begin{cases} (1, 0, \ldots, 0) \\ \vdots \\ (0, \ldots, 0, 1) \\ (p_1 + 1, p_2, p_3, \ldots, p_d) \\ \vdots \\ (p_1, p_2, p_3, \ldots, p_d + 1) \\ (p_1 - 1, p_2, p_3, \ldots, p_d) \\ \vdots \\ (p_1, p_2, p_3, \ldots, p_d - 1) \end{cases}$$

Clearly, each of these sets $DEL(I)$, $SUB(I)$, and $ADD(I)$ of possible moves can be enlarged (or modified) to enforce this algorithm to search the parameter space more aggressively, but an obvious need for this is not seen presently.

Next, we describe how the three basic moves of the D/S/A algorithm can be used to generate index sets $I_k(P_n)$, that seek to minimize the empirical risk function, $f_E(I)$, over all index sets I of size less than or equal to k, $k = 1, \ldots, K_n$ (Box 1).

In our case of the squared error loss function, with full data, the empirical risk function is simply the mean squared error (cf. residual sum of squares) for $\hat{\Psi}_I(P_n)$

$$f_E(I) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\Psi}_I(P_n)(W_i))^2.$$

Denote the best (in terms of empirical risk) index set I of size less than or equal to k, $k = 1, \ldots, K_n$, by

$$I_k^\star(P_n) \equiv \underset{\{I:|I|\leq k,\ I\in\mathcal{I}\}}{\operatorname{argmin}} f_E(I).$$

The D/S/A algorithm, described in Box 1, returns for each k, an index set $I_k(P_n)$ that aims to approximate (or equal) $I_k^\star(P_n)$.

Box 1. Deletion/Substitution/Addition algorithm for optimizing the empirical risk function.

1. **Initialization.** Set $I_0 = \emptyset$ and $BEST(k) = \infty$, $k = 1, 2, \ldots$, where $BEST(k)$ represents the current lowest value of the objective function $f = f_E$ for index sets I of size k. Let $BEST.SET(k)$ represent the actual index sets so that $f(BEST.SET(k)) = BEST(k)$.

2. **Algorithm (*).** Let $k = |I_0|$. Find an optimal updated index set I^- of size $k-1$, among all allowed **deletion** moves: $I^- \equiv \text{argmin}_{I \in DEL(I_0)} f(I)$. If $f(I^-) < BEST(k-1)$, then set $I_0 = I^-$, $BEST(k-1) = f(I^-)$, $BEST.SET(k-1) = I_0$, and go back to (*).

   Otherwise, find an optimal updated index set $I^=$ of the same size k as I_0, among all allowed **substitution** moves: $I^= \equiv \text{argmin}_{I \in SUB(I_0)} f(I)$. If this update improves on I_0, that is, $f(I^=) < f(I_0)$, then set $I_0 = I^=$, $BEST(k) = f(I^=)$, $BEST.SET(k) = I_0$, and go back to (*).

   Otherwise, find an optimal updated index set I^+ of size $k+1$, among all allowed **addition** moves: $I^+ \equiv \text{argmin}_{I \in ADD(I_0)} f(I)$. Set $I_0 = I^+$. If this update improves on I_0, that is, $f(I^+) < f(I_0)$, then set $BEST(k+1) = f(I^+)$, and $BEST.SET(k+1) = I_0$. Go back to (*).

3. **Stopping rule.** Run the algorithm until the current index set size $k = |I_0|$ is larger than a user-supplied max. size or until $f(I^+) - f(I_0) < \Delta$ for a user-specified $\Delta > 0$. Denote the last set I by $I_{\text{final}}(P_n)$.

Note that the D/S/A algorithm is such that $BEST(k)$ is decreasing in k, since addition moves only occur when they result in a decrease in risk over the current index set size. Thus, the best subset of size k is also the best subset of size less than or equal to k. We also note that this algorithm gives priority to moves which make the fit smaller, and it avoids getting "trapped" by always carrying out the addition move.

Unlike previously proposed forward/backward selection approaches, the D/S/A algorithm performs an extensive search of the parameter space, truly aimed at minimizing the empirical risk function over all index sets of a given size.

3.1 Simple example to illustrate D/S/A algorithm.

Consider the regression setting so that $L(O,\psi) = (Y - \psi(W))^2$, and suppose that we parameterize each allowed regression function as linear combinations of tensor products of the polynomial powers. Suppose that $W = (W_1, \ldots, W_4)$ (i.e., $d = 4$) and that the current model (i.e., I_0) in the D/S/A algorithm is given by $Y = W_1W_2W_3 + W_2W_4^5$. Note that the current size is $k = 2$, the corresponding indices are $\vec{p}_1 = (1,1,1,0)$, $\vec{p}_2 = (0,1,0,5)$, and $I_0 = \{\vec{p}_1, \vec{p}_2\}$.

A *deletion* move simply means removing one of the terms of the current model and fitting a model of size $k-1$. Thus, the deletions set, $DEL(I_0)$, contains two index sets of size $k = 1$

$$DEL(I_0) = \Big\{\{\vec{p}_1\}, \{\vec{p}_2\}\Big\} = \Big\{\{(1,1,1,0)\}, \{(0,1,0,5)\}\Big\}.$$

The *substitution* moves involve replacing the j^{th} term for $j = 1, \ldots, k$ with a new term, keeping the size of the model fixed at k. The possible substitution moves are given by:

$$SUB(I_0) = \begin{cases} W_1^2W_2W_3 + W_2W_4^5 & \vec{p}_1 = (2,1,1,0) \\ W_1W_2^2W_3 + W_2W_4^5 & \vec{p}_1 = (1,2,1,0) \\ W_1W_2W_3^2 + W_2W_4^5 & \vec{p}_1 = (1,1,2,0) \\ W_1W_2W_3W_4 + W_2W_4^5 & \vec{p}_1 = (1,1,1,1) \\ W_2W_3 + W_2W_4^5 & \vec{p}_1 = (0,1,1,0) \\ W_1W_3 + W_2W_4^5 & \vec{p}_1 = (1,0,1,0) \\ W_1W_2 + W_2W_4^5 & \vec{p}_1 = (1,1,0,0) \\ W_1W_2W_4^5 + W_1W_2W_3 & \vec{p}_2 = (1,1,0,5) \\ W_2^2W_4^5 + W_1W_2W_3 & \vec{p}_2 = (0,2,0,5) \\ W_2W_3W_4^5 + W_1W_2W_3 & \vec{p}_2 = (0,1,1,5) \\ W_2W_4^6 + W_1W_2W_3 & \vec{p}_2 = (0,1,0,6) \\ W_4^5 + W_1W_2W_3 & \vec{p}_2 = (0,0,0,5) \\ W_2W_4^4 + W_1W_2W_3 & \vec{p}_2 = (0,1,0,4) \end{cases}$$

We want also to note that, if the total number of terms in the tensor products is bounded by $s_2 = 3$, then the substitution move which would not be allowed, $W_1W_2W_3W_4 + W_2W_4^5$, would be replaced by these alternate moves: $W_2W_3W_4 + W_2W_4^5$, $W_1W_3W_4 + W_2W_4^5$, $W_1W_2W_4 + W_2W_4^5$.

If none of these substitution moves improve RSS, then the D/S/A algorithm finds the best fit among the following *addition* moves:

$$ADD(I_0) = \begin{cases} W_1 + W_1W_2W_3 + W_2W_4^5 & \vec{p}_3 = (1,0,0,0) \\ W_2 + W_1W_2W_3 + W_2W_4^5 & \vec{p}_3 = (0,1,0,0) \\ W_3 + W_1W_2W_3 + W_2W_4^5 & \vec{p}_3 = (0,0,1,0) \\ W_4 + W_1W_2W_3 + W_2W_4^5 & \vec{p}_3 = (0,0,0,1) \\ W_1^2W_2W_3 + W_1W_2W_3 + W_2W_4^5 & \vec{p}_3 = (2,1,1,0) \\ W_1W_2^2W_3 + W_1W_2W_3 + W_2W_4^5 & \vec{p}_3 = (1,2,1,0) \\ W_1W_2W_3^2 + W_1W_2W_3 + W_2W_4^5 & \vec{p}_3 = (1,1,2,0) \\ W_1W_2W_3W_4 + W_1W_2W_3 + W_2W_4^5 & \vec{p}_3 = (1,1,1,1) \\ W_2W_3 + W_1W_2W_3 + W_2W_4^5 & \vec{p}_3 = (0,1,1,0) \\ W_1W_3 + W_1W_2W_3 + W_2W_4^5 & \vec{p}_3 = (1,0,1,0) \\ W_1W_2 + W_1W_2W_3 + W_2W_4^5 & \vec{p}_3 = (1,1,0,0) \\ W_1W_2W_4^5 + W_1W_2W_3 + W_2W_4^5 & \vec{p}_3 = (1,1,0,5) \\ W_2^2W_4^5 + W_1W_2W_3 + W_2W_4^5 & \vec{p}_3 = (0,2,0,5) \\ W_2W_3W_4^5 + W_1W_2W_3 + W_2W_4^5 & \vec{p}_3 = (0,1,1,5) \\ W_2W_4^6 + W_1W_2W_3 + W_2W_4^5 & \vec{p}_3 = (0,1,0,6) \\ W_4^5 + W_1W_2W_3 + W_2W_4^5 & \vec{p}_3 = (0,0,0,5) \\ W_2W_4^4 + W_1W_2W_3 + W_2W_4^5 & \vec{p}_3 = (0,1,0,4) \end{cases}$$

3.2 Available Options.

Dimension Reduction. Depending on the application at hand, it can be worthwhile to transform and/or reduce the number of given explanatory variables. To reduce the data, we compute d T-statistics corresponding to the main effects of $W_1, \ldots, W_d$ by fitting d univariate regressions. Next, we rank these statistics, possibly in absolute value, in decreasing order $\hat{R}(1), \ldots, \hat{R}(d) \subset \{1, \ldots, d\}$ yielding our ordered covariates $W_{\hat{R}(1)}, W_{\hat{R}(2)}, \ldots, W_{\hat{R}(d)}$. Then, one can input the set $(W_{\hat{R}(1)}, \ldots, W_{\hat{R}(s_0)})$, of length s_0, as the vector of covariates into the D/S/A algorithm and can choose whether or not to select s_0 via cross-validation. This reduction was done in Section 5.

Derivative-based importance measures. In prediction problems, a common and practical question is to assess the *importance* of a variable, or set of variables, in terms of its predictive ability for the outcome of interest. For instance, in microarray experiments, one is interested in determining how

important each gene (or set of genes) is for the prediction of a particular biological or clinical outcome. Measures of variable importance can assist in the identification of a subset of marker genes for the outcome.

Various measures of importance exist in the literature including loss-function based importance measures (Dudoit et al., 2003), and we will describe the measure that we used in Section 5.

As before, let the data be n observations of (Y, W), where Y is the outcome of interest and W is a d-dimensional vector of covariates for which we would like a measure of importance. Let $h(W) = E(Y|W)$ and let $\alpha(j)$ denote the importance measure for variable W_j:

$$\alpha(j) = E\left(\frac{d}{dW_j}E[h(W)|W_j]\right)$$

To estimate this importance measure, consider a fitted regression model denoted by $\hat{h}(W)$. Based on the idea of counterfactual variables in the causality literature (van der Laan and Robins, 2003), we are getting a sense of the importance of variable W_j for $j = 1, \ldots, d$ by seeing what happens when $W = w$ for a given variable of interest. Given a fit from a particular model $\hat{h}_b(W)$ for $b = 1, \ldots, B$, let $\bar{h}_j(w) = \frac{1}{n}\sum_i \hat{h}_b(W_{1,i}, \ldots, W_{j-1,i}, w, W_{j+1,i}, \ldots, W_{d,i})$. The importance measure, for the following types of variables, can be estimated as follows:

- Continuous
  $\hat{\alpha}_b(j) = \frac{\int_{w\in\mathcal{W}_j} |\frac{d}{dw}\bar{h}_j(w)|dw}{\int_{w\in\mathcal{W}_j} dw}$ where $\mathcal{W}_j$ represents the set of possible values of W_j.
- Binary
  $\hat{\alpha}_b(j) = \mid \bar{h}_j(1) - \bar{h}_j(0) \mid$
- General Discrete
  $\hat{\alpha}_b(j) = \frac{\sum_{l=1}^{K-1} |\bar{h}_j(l+1) - \bar{h}_j(l)|}{K-1}$ where $w \in \{1, \ldots, K\}$.

The final estimate of the importance measure is then a weighted average of $\hat{\alpha}_b(j)$ across many b-specific fits. We can easily accomplish this in two ways. The first approach is to use the fits for all $s \in \mathcal{A}_n$ and estimate $\hat{\alpha}_s(j)$ for all s. The second approach is to reduce the data to $V \subset W$ reasonably

important variables. Then form B random subsets of variables of a specified size from V. The matrix S_b identifies these subsets where each row represents a given subset of variables. Then for a given variable, its importance measure is estimated across fits as:

$$\hat{\alpha}(j) = \frac{\sum_{b=1}^{B} \hat{\alpha}_b(j) I(W_j \in S_b) \mathrm{wt}_b}{\sum_{b=1}^{B} I(W_j \in S_b) \mathrm{wt}_b} \tag{4}$$

In equation 4, wt represents a weight for a particular fit which can be the cross-validated risk of the given regression model for example. For instance, to measure variable importance in Section 5, we reduced the data to the 10 most important variables based on univariate regressions and formed all subsets of size 3 from these 10 variables.

3.3 Other Approaches.

The estimation road-map offered by van der Laan and Dudoit (2003) and van der Laan et al. (2004) inspired the D/S/A algorithm and lists as its three main steps:

1. Definition of the parameter of interest in terms of a loss function.
2. Construction of candidate estimators based on a loss function.
   Define a finite collection of candidate estimators for the parameter of interest based on a sieve of increasing dimension approximating the complete parameter space. For each element of the sieve, the candidate estimator is chosen as the minimizer of the empirical risk based on the observed data loss function.
3. Cross-validation for estimator selection and performance assessment.

The D/S/A algorithm incorporates Steps 2 and 3 by first forming candidate estimators as outlined in step two. It then uses cross-validation to select the optimal estimator among the candidates it formed.

There are many approaches to regression problems in the statistics and machine learning literature which can produce candidate estimators (Breiman et al. (1984),Friedman (1991), Ruczinski et al. (2003), Efron et al. (2004)). The results for cross-validation given by van der Laan and Dudoit (2003) are

very general, and hence they apply to any set of candidate estimators. The D/S/A algorithm does not differ in its use of cross-validation but in the way it produces candidate estimators.

Friedman (1991) introduced MARS, an adaptive procedure for regression, which uses linear splines as basis functions. Stone et al. (1997) developed a hybrid of MARS. Support Vector Machines (SVMs) have been well-promoted for classification and have been adapted for regression, sometimes referred to as support vector regression (Shawe-Taylor and Cristianini, 2004). To use the SVM, one needs to define a kernel and its corresponding parameters which requires time spent fine-tuning. Logic Regression (Ruczinski et al., 2003) is an adaptive regression method and performs a very thorough search by allowing many different moves in its tree growing process. It works in a specialized context, to construct predictors as Boolean combinations of binary covariates.

However, many existing methods for constructing candidate estimators are not aggressive enough for the types of datasets encountered in genomics. They either only accomodate variable main effects or are too rigid to generate a good set of candidate estimators. These approaches do not aim to minimize the empirical mean of a loss function over specified subspaces of the complete parameter space. Instead, they rely on forward/backward-like local optimization steps. For example, while regression trees allow interactions among variables, the candidate tree estimators are generated according to a limited set of moves, amounting to forward selection (node splitting) followed by backward elimination (tree pruning).

The D/S/A algorithm keeps running because it can always improve the objective function by adding a term or in some cases by altering an existing term. Once it adds a term, it then has the ability to try to delete a term or alter a term and can avoid getting "stuck." It performs a very aggressive search. Logic Regression also performs an aggressive search. Yet, many other methods do not perform thorough enough searches.

In the next section, we compare our approach to Logic Regression, MARS, and SVMs. Two implementations of the D/S/A algorithm for histogram regression (Molinaro and van der Laan, 2004) and neural networks (Durbin and Dudoit, 2004) are being developed.

4 Simulations.

In this section, we will present the results of applying the D/S/A algorithm to simulated data sets. All simulations and data analyses were done using machines available in the labs of the Statistical Computing Facility (University of California, Berkeley). These are Sun workstations which have UltraSparc II processors ranging in speed from around 200 MHz to 440 MHz. They typically have 128 MB to 256 MB of RAM.

We conducted a variety of simulations to see how the algorithm performs in different settings. The first set of simulations use the D/S/A algorithm in its simplest form while later simulations impose different constraints on the algorithm. In all simulations, no additional constraints have been placed on β nor did we reduce the data. Dimension reduction is done for the data analysis (Section 5).

The D/S/A algorithm first is implemented without constraints (brakes); we are not restricting the number of tensor products, k, and thus not using cross-validation to select k. We run the algorithm until we reach a k that gives a minimal residual sum of squared error (RSS). Next, the algorithm is implemented with a cross-validated constraint placed on the number of tensor products in the regrssion, i.e., selecting k via v-fold cross-validation. The cross-validated D/S/A algorithm is then compared to: the R function `stepAIC()`, forward selection with cross-validation (*fscv*), Logic Regression, and Multivariate Adaptive Regression Splines (MARS). Finally, the algorithm which places brakes on the complexity of each tensor product: $m_2(I) = \max_{\vec{p} \in I} \sum_{j=1}^{d} \mathrm{I}(p_j \neq 0)$; $m_2(I) \leq s_2$ and $m_3(I) = \max_{\vec{p} \in I} \sum_{j=1}^{d} p_j$; $m_3(I) \leq s_3$, and thereby incorporates the *alternate-substitution* moves is implemented and used in the Logic Regression and MARS comparisons.

In each of the simulations, an $n \times d$ covariate matrix, W, is generated from a given probability distribution, e.g. normal, uniform, Bernoulli. The true mean linear polynomial regression model, $E(Y|W)$, is either manually or randomly generated. The outcome Y is then generated from the true mean linear polynomial regression model with no noise or a Gaussian noise with mean 0 and standard deviation σ. The D/S/A algorithm is then used to minimize $f_{E,s}(I)$, and the procedure may be repeated a number of times.

4.1 Implementation without constraints.

The first set of simulations explore the performance of the D/S/A algorithm itself. Therefore, cross-validation is not yet employed, and the D/S/A algorithm is run on the *learning* set without using cross-validation to select the size k.

The purpose of these simulations is to establish to what degree the D/S/A algorithm is truly capable of finding the global minimum (i.e., the optimal predictor $W \to \psi_0(W) = E_0(Y|W)$) when n is large enough. In the following simulations, the true regression model is randomly generated (see Sinisi and van der Laan (2004)).

Numerical results obtained from this simulation are available in (Sinisi and van der Laan, 2004). We found that the algorithm succeeded in minimizing $f_{E,s}(I)$; both sensitivity and specificity is 100% indicating that the algorithm is successful in fitting true *simple* regressions in the case of zero error which corresponds to choosing a very large sample size. These results are encouraging and led to further exploration of the algorithm's capabilities.

The next step is to see what happens when some noise is added to these regression models. We used a normal distribution with mean 1 and standard deviation 0.5 to generate $W_{n,d}$ for $E_5[Y|W]$ and $E_6[Y|W]$. The outcome Y is generated from the randomly chosen true regression model with zero error or Gaussian error with mean 0 and standard deviation 1. These were run only once.

The following two models were generated, first with $\varepsilon = 0$ and then with $\varepsilon \sim \mathcal{N}(0,1)$.

$$E_5[Y|W] = W_0 W_1^2 W_2^2 + W_0 W_1 W_2^2 W_3 + W_2^3 + W_4^4$$

$$E_6[Y|W] = W_0 + W_0 W_{49} W_{99} + W_{24}^5 + W_{17} W_{30} W_{53} W_{62} W_{78} W_{88}$$

Table 1 displays the results of this simulation which compares two models with zero error and a Gaussian error. The truth was identified in all cases where $\varepsilon = 0$ or $\varepsilon \sim \mathcal{N}(0,1)$. The number of moves the algorithm needed to make in order to converge are displayed as well for this simulation. Based on Table 1, a large number of covariates does not affect the convergence rate since the number of moves performed when $d = 100$ is less than the number of moves performed when $d = 5$.

$E[Y\|W]$	n	d	*sens*	*spec*	RSS_n	RSS_0	moves	subs	adds	dels
$E_5[Y\|W]$	1000	5	100%	100%	0.0000	0.0000	30	22	6	2
$E_5[Y\|W]^*$	1000	5	100%	80%	1.074	1.080	30	23	6	1
$E_6[Y\|W]$	1000	100	100%	100%	0.0000	0.0000	21	17	4	0
$E_6[Y\|W]^*$	1000	100	100%	100%	0.9576	0.9572	21	17	4	0

Table 1: **DSA unconstrained.** Comparing $\varepsilon = 0$ and $\varepsilon \sim \mathcal{N}(0,1)^*$. *sens*: sensitivity, *spec*: specificity, RSS_n: $RSS/(n-b)$ represents the estimate of the variance of the error where b is the number of independent variables in fitted model, RSS_0: true variance of the error, *moves*: number of moves made by the algorithm, *subs*: number of substitution moves made, *adds*: number of addition moves made, *dels*: number of deletion moves made, *: indicates the model for which $\varepsilon \sim \mathcal{N}(0,1)$.

4.2 D/S/A algorithm with cross-validated size versus the `stepAIC()` function in R.

The next set of simulations address the performance of cross-validation in making sure that the algorithm does not select too many variables, and thereby over-fits, by comparing it to the R function `stepAIC()`. We first generated the covariate matrix $W_{n,d}$ of n i.i.d. observations of d variables W_i, $i = 1, \ldots, d$ from a uniformly distributed distribution between 1 and 10. Then, we manually generated the following three true regression models:

$$E_1[Y|W] = W_1 + W_2^2$$

$$E_2[Y|W] = W_1 W_3$$

$$E_3[Y|W] = W_1 W_3 + W_5^2 + W_7 W_{10}$$

Using the models, we next generated the outcome Y with a Gaussian error with mean 0 and standard deviation 1. Then ran the procedure once with the cross-validated D/S/A algorithm (*DSA1-CV*) and `stepAIC` and reported the final size of the fitted model chosen by each method, $\hat{k}$, and an estimate of the true risk, $\hat{r}$.

The D/S/A algorithm creates variables data-adaptively and therefore does not require enumeration of all potential variables. `StepAIC` does require enumeration of all variables. To compare the two black-box algorithms

(data → predictor), we enumerated all main terms, squared terms, and pairwise interactions. This is a preliminary simulation to compare the D/S/A algorithm with a cross-validated constraint on the size of the model with the forward selection algorithm (enumerating all terms) using AIC to select the size of the model. (In Section 4.3, we will compare our *DSA1-CV* algorithm to forward selection with cross-validation.) For this simulation, we are interested in whether or not each method fits the true model and the true risk of the selected model. The true risk is estimated by setting aside a large sample of independent observations, a *test set*, and calculating the risk based on the fitted model on this set of observations. In this particular case, a test set of size 20,000 was used to estimate the true risk.

In the first simulation (row 1, table 2), both our method and `stepAIC` selected the exact true model. However, in the next two simulations (rows 2-3, table 2), our method fitted the truth exactly while `stepAIC` heavily overfitted the model (col 4, table 2). AIC's tendency to over-fit is well-known in the statistical literature, but the over-fitting did not hurt the risk estimate because the estimated risk from the fitted model produced by `stepAIC` is nearly the same as the estimated risk given by our method's fitted model.

$E[Y\|W]$	n	d	$\hat{k}_{AIC}$	$\hat{k}_{DSA1-CV}$	$\hat{r}_{AIC}$	$\hat{r}_{DSA1-CV}$
$E_1[Y\|W]$	5000	3	2	2	0.9963	0.9963
$E_2[Y\|W]$	5000	10	19	1	0.9995	0.9932
$E_3[Y\|W]$	5000	10	22	3	1.0174	1.0106

Table 2: Comparing `stepAIC` to DSA-CV algorithm with cross-validated constraint on size under 2-fold cross-validation. $\hat{k}$: size of the final fitted model for each method, $\hat{r}$: estimate of the true risk, based on 20,000 independent observations, of the final model chosen by both methods.

4.3 Comparison to forward selection with cross-validation.

This simulation study (Table 3) compares the D/S/A algorithm (*DSA1-CV* to a type of forward selection with cross-validation (fscv) algorithm. The forward selection algorithm makes all the same *addition* moves as our method but does not carry out the deletion and substitution moves.

The true model, (Table 3), is $y = 4w + 3w^3 - 2w^5 + \varepsilon$, where $w \sim U(1, 5)$, ε is normal with mean 0 and standard deviation 1, and $y \in (-6000, 8)$. Both

methods were run under 2, 5, and 10-fold cross-validation with the maximum number of terms in the model pre-set at 10.

The fscv algorithm, naturally, picks a model with about 5 or more terms, not having the deletion or substitution step to get rid of terms involving even powers, and thus has a sensitivity of 100% in all cases. *DSA1-CV* picks a smaller model on average with a higher specificity as expected. The risk estimates are approximately the same for both methods. The D/S/A algorithm seems to be more efficient for smaller sample sizes than the fscv algorithm.

4.4 Logic Regression.

Logic Regression (Ruczinski et al., 2003) is a very useful regression method currently available, and it can handle a variety of problems including linear regression, logistic regression and classification and can be extended to other problems by defining an appropriate score function. Both the D/S/A algorithm and Logic Regression is an adaptive regression methodology that attempts to construct predictors. However, the goal of Logic Regression is to find predictors that are Boolean (logical) expressions, and thus is applied when the covariates in the data to be analyzed are primarily binary. The D/S/A algorithm can handle any combination of continuous and discrete covariates. It is important to compare the two methods when applied to binary variables, and this simulation is an initial attempt at comparing the two. Logic Regression uses a cross-validated constraint on the complexity of each tree, which corresponds to the complexity of our tensor products (implemented by *DSA2-CV*, Table 4). Thus, Logic Regression is compared to two implementations of the D/S/A algorithm, (*DSA1-CV* and *DSA2-CV*) referred to as dsa1 and dsa2, respectively, in Table 4.

DSA1-CV has been described previously; it uses cross-validation to select k, the number of tensor products. *DSA2-CV* uses cross-validation to select the number of tensor products and places a brake on the complexity of each tensor product thereby involving the alternate-substitution moves. Specifically, we are limiting the order of interactions to be no greater than a specified value, $m_2(I) = \max_{\vec{p} \in I} \sum_{j=1}^{d} I(p_j \neq 0) \leq s_2$. In the case of binary covariates, $m_2(I) \equiv m_3(I)$ and thus we did not need to select s_3 via cross-validation.

The true model was generated from $y = \beta_1(w_1 w_3(1 - w_2)) + \beta_2((1 - w_1)w_3(1-w_2)) + \beta_3(w_7 w_{10}) + er$, where $w_i \sim \mathcal{B}(0.7)$, $1 \leq i \leq 10$, $\beta \sim \mathcal{N}(1, 1)$,

Table 3: *Simulation study* **FSCV Comparison**. Data simulated from $y = 4w + 3w^3 - 2w^5 + er$, where $w \sim U(1,5)$ and $er \sim N(0,1)$. Candidate estimator was chosen over 50 repetitions of three sample sizes (col 1), three v-fold cross-validations (col 2) for all algorithms (col 3). The results (cols 4-9) in the table are based on an independent test sample of $n = 10000$. col 4 is the average of the 50 risks for each method, col 5 is the standard deviation of the risks over the 50 reps, col 6 is the average size (number of basis functions), col 7 is the sensitivity, col 8 is the specificity, and col 9 is the ratio of averaged risks (col 4) $-$ optimal risk, (ours/fscv).

Sample			50 Repetitions					
Size	$v-fold$	Method	mean	std dev	avg size	sens	spec	ratio
250	2	ours	1.043	.018	3.62	83%	71%	1
		fscv	1.045	.019	5.36	100%	57%	.950
	5	ours	1.044	.018	3.64	82%	71%	1
		fscv	1.046	.019	5.46	100%	56%	.960
	10	ours	1.043	.018	3.62	82%	72%	1
		fscv	1.046	.019	5.52	100%	55%	.939
500	2	ours	1.031	.006	3.24	91%	86%	1
		fscv	1.031	.007	5.28	100%	57%	.980
	5	ours	1.030	.006	3.30	89%	84%	1
		fscv	1.031	.007	5.30	100%	57%	.965
	10	ours	1.030	.006	3.34	89%	83%	1
		fscv	1.031	.007	5.34	100%	57%	.967
1000	2	ours	1.026	.005	3.54	85%	75%	1
		fscv	1.026	.005	5.28	100%	57%	1.014
	5	ours	1.027	.005	3.46	84%	75%	1
		fscv	1.026	.005	5.18	100%	58%	1.023
	10	ours	1.026	.005	3.44	85%	77%	1
		fscv	1.026	.005	5.28	100%	57%	1.015

and $er \sim N(0,1)$. It has been pointed out that the model can be reduced to:

$$\beta_1(w_3(1-w_2)) + (\beta_2-\beta_1)(w_3(1-w_1)(1-w_2)) + \beta_3(w_7w_{10})$$

or

$$\beta_2(w_3(1-w_2)) + (\beta_2+\beta_1)(w_3w_1(1-w_2)) + \beta_3(w_7w_{10}).$$

Thus, the true number of leaves is 7.

When running Logic Regression, the preset maximum number of allowed trees matched the number of terms in the true model. The results of both Logic Regression and DSA-CV depend on the fine tuning parameters such as number of folds, number of trees or maximum number of tensor products, and number of leaves or tensor product complexity measure. In terms of prediction, this simulation shows that the D/S/A algorithm is competitive with Logic Regression since the risk ratios are roughly one. When using binary covariates, both methods produce models that are easy to interpret. Logic Regression provides the user with enhanced interpretability by using intuitive operators. See Sinisi and van der Laan (2004) for two other comparisons to Logic Regression.

4.5 Multivariate Adaptive Regression Splines.

Multivariate Adaptive Regression Splines (MARS) (Friedman, 1991) is a method for flexible regression modeling of high-dimensional data. It can be viewed as a generalization of stepwise linear regression or a modication of the CART (Breiman et al., 1984) method. MARS uses expansions in the form of linear splines.

To compare the D/S/A algorithm to MARS, we used the `mars{mda}` `R` function to run MARS. We used the model taken from Section 4.3 of the MARS paper (Friedman, 1991):

$$y = 10\sin(\pi w_1 w_2) + 20(w_3 - \frac{1}{2})^2 + 10w_4 + 5w_5 + er,$$

where $w_i \sim \mathcal{U}(0,1)$, $1 \leq i \leq 10$ and $er \sim N(0,1)$.

Both methods allow the user to set an optional integer specifying the maximum interaction degree (`degree`) and number of model terms (`nk`). To compare both methods under the same level of constraint, we set `nk` to ten and `degree` to two. For the D/S/A algorithm, we set the maximum sum of powers to be 5 and used 10-fold cross-validation. The results for

Table 4: *Simulation study* **Logic Regression Comparison**. Data simulated from $y = \beta_1(w_1 w_3(1-w_2)) + \beta_2((1-w_1)w_3(1-w_2)) + \beta_3(w_7 w_{10}) + er$, where $w_i \sim \mathcal{B}(0.7)$, $1 \leq i \leq 10$, $\beta \sim \mathcal{N}(1,1)$, and $er \sim N(0,1)$. Candidate estimator was chosen over 10 repetitions of two sample sizes (col 1), three v-fold cross-validations (col 2) for both our algorithm and logic regression (col 3). The results (cols 4-7) in the table are based on an independent test sample of $n = 10000$. col 4 is the average size (number of basis functions for ours and number of leaves for logic), col 5 is the average of the 10 risks (with the L_2 loss function) for each method, col 6 is the standard deviation of the risks over the 10 reps, and col 7 is the ratio of averaged risks (col 5) − optimal risk, (dsa/logic).

Sample			10 Repetitions			
Size	$v-fold$	Method	avg size	mean risk	std dev	ratio
250	2	dsa1	4.1	4.647	.248	.995
		dsa2	4.0	4.636	.254	.992
		logic	5.6	4.664	.255	1
	5	dsa1	4.8	4.705	.319	1.013
		dsa2	4.7	4.723	.297	1.018
		logic	6.0	4.657	.246	1
	10	dsa1	4.8	4.705	.319	1.014
		dsa2	4.7	4.723	.297	1.019
		logic	6.0	4.654	.278	1
1000	2	dsa1	5.0	4.738	.101	1.001
		dsa2	4.9	4.731	.105	.999
		logic	7.0	4.734	.105	1
	5	dsa1	5.0	4.738	.101	1.001
		dsa2	5.0	4.743	.101	1.002
		logic	7.0	4.734	.105	1
	10	dsa1	5.0	4.738	.101	1.001
		dsa2	5.0	4.743	.101	1.002
		logic	7.0	4.734	.105	1

Table 5: *Simulation study* **I. MARS Comparison**. Data simulated from $y = 10\sin(\pi w_1 w_2) + 20(w_3 - \frac{1}{2})^2 + 10w_4 + 5w_5 + er$, where $w_i \sim \mathcal{U}(0,1)$, $1 \leq i \leq 10$ and $er \sim N(0,1)$. Candidate estimator was chosen over 50 repetitions of two sample sizes (col 1) for our algorithm versus MARS (col 2). The results (cols 3-6) in the table are based on an independent test sample of $n = 10000$. col 3 is the average size (number of basis functions for ours and MARS), col 4 is the average of the 50 risks (with the L_2 loss function) for each method, col 5 is the standard deviation of the risks over the 50 reps, and col 6 is the ratio of averaged risks (col 5) − optimal risk, (dsa/mars).

Sample		50 Repetitions			
Size	Method	avg size	mean risk	std dev	ratio
250	dsa (10)	9.50	2.927	1.6	1
	mars (10)	6.38	5.531	.19	0.43
	dsa (15)				
	mars (15)				
500	dsa (10)	9.36	2.511	1.5	1
	mars (10)	6.34	5.333	.12	0.35
	dsa (15)				
	mars (15)				

50 repetitions comparing the estimated risks of each method based on an independent test set of 10,000 are displayed in Table 5. In this setting, it looks as if the D/S/A algorithm outperforms MARS. However, MARS produces more consistent results (smaller risk variance). The D/S/A algorithm has the ability to produce models with very low risk, but it is unable to do this for every repetition under the imposed constraints. Furthermore, if we ran `mars` without specifying `nk`, it tends to pick models of size 14-15 with a much lower risk estimate. Using degree 2 or degree 10 did not radically alter the results.

EDIT once new results are in!

Table 6: *Simulation study* **II. MARS Comparison**. Data simulated from $y = 10\sin(\pi w_1 w_2) + 20(w_3 - \frac{1}{2})^2 + INSERT + er$, where $w_i \sim \mathcal{U}(0,1)$, $1 \leq i \leq 10$ and $er \sim N(0,1)$. Candidate estimator was chosen over 50 repetitions of two sample sizes (col 1) for our algorithm versus MARS (col 2). The results (cols 3-6) in the table are based on an independent test sample of $n = 10000$. col 3 is the average size (number of basis functions for ours and MARS), col 4 is the average of the 50 risks (with the L_2 loss function) for each method, col 5 is the standard deviation of the risks over the 50 reps, and col 6 is the ratio of averaged risks (col 5) − optimal risk, (dsa/mars).

Sample		50 Repetitions			
Size	Method	avg size	mean risk	std dev	ratio
250	dsa (10)				1
	mars (10)				
	dsa (15)				
	mars (15)				
500	dsa (10)				
	mars (10)				
	dsa (15)				
	mars (15)				

5 Data Analysis.

An important problem in contemporary biology is transcription factor binding site identification. The activities of hundreds of sequence specific DNA binding proteins, transcription factors (TFs), play an important role in transcriptional regulation of eukaryotes. TFs are proteins, needed to initiate the transcription of a gene, that bind to regions in the vicinity of genes and as a result regulate the activities of the genes. Each TF, or group of closely related factors, recognizes a unique grouping of short sequence elements, usually between five and fifteen basepairs in length. Identification of these sites is a crucial problem as understanding the components of regulation is a step toward understanding how genes are expressed at all times in the cell life. In this section, we look at the identification of biologically significant transcription factor binding sites in the genome of the yeast *Saccharomyces*

cerevisiae.

This biological problem has been put by Keleş et al. (2002) into a statistical framework by formulating it as a model selection problem. Keleş et al. (2002) model gene expression as a function of short oligonucleotides that represent potential binding sites and use length five motifs, or pentamers, as an initial set of covariates, adopting a stepwise cross-validation methodology with forward selection and backward deletion to choose the most predictive pentamers.

5.1 Cell Cycle Data.

The eukaryotic cell cycle consists of four phases: M (mitosis), S (synthesis, DNA is replicated), G_1, and G_2. During the first gap phase, G_1, cells increase in size, produce RNA, and synthesize protein, and there is a checkpoint ensuring that everything is ready for DNA synthesis. During the gap between DNA synthesis and mitosis, G_2, the cell will continue to grow, produce new proteins, and determine if the cell can proceed to enter mitosis and divide.

Cho et al. (1998) gathered data by using Affymetrix oligonucleotide microarrays to query the abundances of 6,220 mRNA species in synchronized *Saccharomyces cerevisiae* batch cultures. Cells were collected at 17 time points taken at 10 minute intervals to cover nearly two full cell cycles. The time course was divided into early G_1, late G_1, S, G_2, and M phases (G_1-S for replication, S-G_2 for organization of centrosome, and M phase for budding and cell polarity).

For this data analysis, we used 15 of the 17 time points. Time points 90 and 100 minutes were excluded due to the less efficient labeling of their mRNA during the original chip hybridizations (Tavazoie et al., 1999). The outcome was the normalized expression profiles of the most variable 3,000 ORFs. With the 15 time points, we constructed a 3,000 by 15 outcome data matrix.

In yeast, regulatory elements are found almost exclusively upstream from the promoter. There are several known upstream regulatory sequences involved in cell cycle-dependent transcription including the late G_1 elements MCB (MIuI cell cycle box) and SCB (Swi4/6 cell cycle box) and the early G_1 element ECB (early cell cycle box) (Cho et al., 1998). The SCB element has been identified as a regulatory sequence located upstream of genes transcribed in late G_1 and early S. SCB is bound by the SBF transcription factor, a complex of Swi4p and Swi6p (Wolfsberg et al., 1999). The Hap complex

is formed by four proteins (Hap2p, Hap3p, Hap4p, and Hap5p); Hap2p and Hap3p have been shown to bind DNA, and Hap4p acts as activation domain for the complex. The Hap2p-Hap3p binding site contains the conserved motif $CCAAT/C$ (van Helden et al., 1998).

Upstream regions of all genes should be searched for other known yeast regulatory sequences, such as the ABF1 and RAP1 transcription factor binding sites, the stress response element STRE with consensus sequence AGGGG, and the SFF factor which acts with MCM1 to control cell cycle regulated genes. Other cell cycle period-specific transcription factors such as Swi5 and MCM1 do not have a highly conserved binding sequence, making it difficult to search genomic sequence for possible action sites accurately.

5.2 Applying the D/S/A algorithm.

Previous work on cell cycle regulation in yeast suggests that more than one sequence element may be responsible for transcription at the same phase (e.g., SCB and MCB both regulate late G_1 mRNA expression). Wolfsberg et al. (1999) predicts that a variety of elements can be responsible for transcription at each phase of the cell cycle. It is important to look at interactions between motifs, and the D/S/A algorithm is one adaptive regression approach that can easily search through two-way and multi-way interactions of explanatory variables.

As many transcription factors bind to short, highly conserved stretches of DNA, many analyses focuses on short oligomers of length five or six, pentamers or hexamers. We focus on pentamers for comparison to the results of Keleş et al. (2002). After having retrieved the set of upstream sequences from the regulatory family, the number of occurrences of all oligonucleotides of the selected size, five in our case, are counted. There are 512 distinct pairs of pentamers and reverse complements. For these 512 motifs, we form sequence motif scores which represent the proportion of occurences of the given motif. We treat the sequence motif scores (which is a proportion between zero and one) as explanatory variables, and model gene expression as a function of these pentamers present in presumptive transcription control regions. The D/S/A algorithm is used to extract the pentamers that are most relevant.

The D/S/A algorithm was ran three different ways when analyzing the data: (1) select the number of tensor products, s_1, via cross-validation; (2) select the number of tensor products, s_1, and two complexity measures of the tensor products, s_2, s_3, via cross-validation where $m_2(I) = \max \sum_{j=1}^{d} \mathrm{I}(p_j \neq$

0) $\leq s_2$ and $m_3(I) = \max \sum_{j=1}^{d} p_j \leq s_3$; or (3) select s_0, s_1, s_2, and s_3 via cross-validation, where s_0 represents the dimension of the vector of covariates and can range between 1 and 512. Each implementation ran with a maximum size model of 5 under 2-fold cross-validation.

The data reduction done in the third implementation ranks the main effects of the 512 given covariates based on the training set. An average of 29 (or fewer pentamers) have significant main effects for $p \leq 0.05$ across the 15 time points while an average of only 8 pentamers have significant main effects at the 0.01 significant p-level. In this particular data set, the reduction steps are essential for reducing the noise by narrowing the candidate covariates down to a significant set.

The variable importance measures calculated using equation 4, where $\hat{\alpha}_b(j)$ is estimated for continuous variables, for each time point are given by Table 8 (displayed for the first four time points). To calculate these measures we reduced the data to the top 10 variables (ranked by $\mid T \mid$) and formed all subsets of 3 variables from these 10 at each time point. Given that the main effects of a limited number of pentamers on average was small relative to the total number of pentamers (29 for $p \leq 0.05$, 8 for $p \leq 0.01$), we chose to form importance measures on the ten most significant pentamers for each time point.

Summary of Previous Results. Keleş et al. (2002) model gene expression on sequence motif scores using a forward and backward stepwise selection method embedded in Monte Carlo cross-validation allowing for main effects and two-way interactions. The scores they use as explanatory variables incorporate the number of occurences of the motifs and their positions with respect to the gene's translation start site. They begin with pentamers as sequence motifs.

Keleş et al. (2002) report *experiment specific importance measures*, $R_w(n)$, on selected pentamers for the first four time points (0, 10, 20, and 30 minutes). The final model given by their feature selection method is not shown. $R_w(n)$ represents a rank weighted proportion of the number of times that motif w is selected within the total number of splits, selected at random from their method. If a motif has $R_w(n) = 1$, then it entered the model first in all of the splits. A motif that is never selected will have an $R_w(n) = 0$. The three pentamers with the highest $R_w(n)$ are (Keleş et al., 2002):

- T = 0 minutes
  AGGGG/CCCCT [stre]

ACGCG/CGCGT [mcb]
GAAAA/TTTTC [ecb]

- T = 10 minutes
  AAACA/TGTTT [ste12]
  CTTAA/TTAAG
  GTTTA/TAAAC [sff]

- T = 20 minutes
  ACGCG/CGCGT [mcb]
  CGCGA/TCGCG [scb]
  AGGGG/CCCCT [stre]

- T = 30 minutes
  ACGCG/CGCGT [mcb]
  CCACA/TGTGG
  CGCGA/TCGCG [scb]

MCB has a measure of one at 20 and 30 minutes.

We restricted our analysis to pentamers for direct comparison to the results of (Keleş et al., 2002), however it is of interest to consider longer motifs. This could be done by incorporating the extension method (Keleş et al., 2002) into the D/S/A algorithm for example.

5.3 Results.

Results given by the three implementations of the D/S/A algorithm for the first time point are given in Table 7. Looking at the results for $T = 0$ minutes, the first two reported models (DSA1: $\hat{s}_1 = 5$; DSA2: $\hat{s}_1 = 5, \hat{s}_2 = 3, \hat{s}_3 = 3$) are identical and composed of five terms: a main effect involving the pentamer *AGGGG* and/or its reverse complement (N.B. a pentamer can refer to itself and/or its reverse complement), a two-way interaction of *ACGCG* with *CGAAA*, and three three-way interactions. Pentamers have partial or exact matches to the regulatory elements shown in brackets after the pentamer. The third reported model reduced the data to 55 covariates and produced a similar, yet simplified, model. The two models contain STRE, MCB, and ECB which are the most highly ranked motifs at $T = 0$ based on the work of Keleş et al. (2002). The method was able to select biologically

Table 7: *Yeast Data Analysis* **DSA1-CV, DSA2-CV, DSA3-CV**, 2-fold cross-validation, applied to yeast cell cycle data of (Cho et al., 1998), first time point.

T=0 min
$(s_0 = 512,\ \hat{s}_1 = 5)$ $(AGGGG[stre]) \quad + \quad (ACGCG[mcb])(CGAAA)$ $+ \quad (ATCCC)(CCTTA)(GCAAA) \quad + \quad (AAAAT)(CATCG)(GATGA)$ $+ \quad (ACCCG)(AGGGG[stre])(GAAAA[ecb])$
$(s_0 = 512,\ \hat{s}_1 = 5,\ \hat{s}_2 = 3,\ \hat{s}_3 = 3)$ $(AGGGG[stre]) \quad + \quad (ACGCG[mcb])(CGAAA)$ $+ \quad (ATCCC)(CCTTA)(GCAAA) \quad + \quad (AAAAT)(CATCG)(GATGA)$ $+ \quad (ACCCG)(AGGGG[stre])(GAAAA[ecb])$
$(\hat{s}_0 = 55,\ \hat{s}_1 = 5,\ \hat{s}_2 = 3,\ \hat{s}_3 = 3)$ $(AGGGG[stre]) \quad + \quad (ACGCG[mcb])$ $+ \quad (ATCCC)(CCTTA) \quad + \quad (CATCG)$ $+ \quad (ACCCG)(AGGGG[stre])(GAAAA[ecb])$

relevant pentamers, and perhaps, other identified pentamers play a role in predicting gene expression.

Importance measures for the ten pentamers with the highest ranked univariate T-statistic at the first four time points are displayed in Table 8. $ACGCG/CGCGT$ [MCB], $CGCGA/TCGCG$ [SCB], and $AGGGG/CCCCT$ [STRE] have the highest overall importance measure. This coincides with what Keleş et al. (2002) found. In late G_1, about 20 and 30 minutes, MCB has the highest importance measure and SCB has the second highest importance measure, as expected. ECB is known to be relevant at early G_1. It was selected at 0 minutes from the D/S/A algorithm, but it did not make the top ten variables.

Wolfsberg et al. (1999) identified pentamers and hexamers as potential regulatory motifs by analyzing UCRs of the genes that might have been involved in the cell-cycle dependent transcription regulation. Comparing their significant pentamers ($p \leq 0.05$) to ours, seven pentamers listed at 20 minutes and six pentamers listed at 30 minutes correspond to the seven significant pentamers they listed for late G_1. Keleş et al. (2002) found that most of the late G_1 pentamers identified by Wolfsberg et al. (1999) were picked up be their method as well. However, their results disagreed at other phases. The pentamers that we selected at different phases also differ from the results of Wolfsberg et al. (1999). In conclusion, it seems that the D/S/A algorithm worked reasonably well as a basis for this analysis.

Table 8: *Yeast Data Analysis* **Variable Importance Measures**

$T = 0$ min.		
Index	Pentamer	VIM
157	AGGGG CCCCT [stre]	11.54
42	AAGGG CCCTT	6.71
98	ACGCG CGCGT [mcb]	8.84
310	CCTTA TAAGG	5.23
82	ACCCC GGGGT	7.62
424	GCCCC GGGGC	7.41
192	ATCCC GGGAT	7.24
264	CATCG CGATG	8.11
455	GGGGA TCCCC	4.52
328	CGCGA TCGCG [scb]	7.43

$T = 10$ min.		
Index	Pentamer	VIM
4	AAACA TGTTT [ste12]	2.81
16	AACAA TTGTT [sff]	1.02
17	AACAC GTGTT	2.75
455	GGGGA TCCCC	3.86
157	AGGGG CCCCT [stre]	2.94
329	CGCGC GCGCG	1.95
318	CGAGA TCTCG	2.38
72	ACAGG CCTGT	3.57
254	CAGGA TCCTG	2.12
479	GTTTA TAAAC [sff]	1.16

$T = 20$ min.		
Index	Pentamer	VIM
98	ACGCG CGCGT [mcb]	21.55
328	CGCGA TCGCG [scb]	17.49
397	GACGC GCGTC	9.20
25	AACGC GCGTT	5.24
428	GCGAA TTCGC [scb]	2.43
157	AGGGG CCCCT [stre]	7.19
42	AAGGG CCCTT	4.89
312	CGAAA TTTCG [scb]	1.30
242	CACGA TCGTG	0.97
433	GCGTA TACGC	2.33

$T = 30$ min.		
Index	Pentamer	VIM
98	ACGCG CGCGT [mcb]	10.51
328	CGCGA TCGCG [scb]	7.02
25	AACGC GCGTT	4.23
397	GACGC GCGTC	3.69
433	GCGTA TACGC	2.60
273	CCACA TGTGG	2.07
238	CACAG CTGTG	1.63
282	CCCAC GTGGG	0.02
428	GCGAA TTCGC [scb]	0.95
362	CTCCA TGGAG	2.15

6 Discussion.

The D/S/A algorithm has been developed as a general tool for loss-based estimation inspired by the theoretical results of van der Laan et al. (2004). It is completely defined by the following choices: the loss function; the basis functions defining the parameterization of the parameter space; and the sets of deletion, substitution, and addition moves. As a result, by choosing the appropriate loss function, it can deal with problems such as multivariate prediction and density/hazard estimation. The D/S/A algorithm, as presented in this article, has been implemented in the context of polynomial regression for the prediction of a univariate outcome. The current implementation of the D/S/A algorithm has the following options available: (1) to choose the number of basis functions by v-fold cross-validation, (2) the option to reduce the data based on univariate regressions to have no more than s_0 candidate covariates where s_0 can be chosen via cross-validation, (3) the option to restrict the order of interaction of candidate tensor products to be no higher than a specified limit s_2 and to choose s_2 also with cross-validation, (4) the option to restrict the sum of polynomial powers of candidate tensor products to be no higher than a specified limit s_3 and to choose s_3 again with cross-validation, and (5) the option to report variable importance measures.

Many regression procedures exist including Logic Regression (Ruczinski et al., 2003) and MARS (Friedman, 1991). Logic Regression is a method freely available in `R` to find interactions between binary inputs associated with an output. It introduced the idea of *substitutions* by defining a number of permissible moves in its tree growing process. It can be adapted to other statistical problems by using the appropriate score function. When faced with binary explanatory variables, Logic Regression is a nice tool to use both for its predictive capabilities and its ease of interpretation. Barron and Xiao argue in favor of their multivariate adaptive polynomial synthesis (MAPS) method over MARS (Friedman, 1991, pg. 67-82). At the time of writing their discussion, Barron had only implemented the forward stepwise synthesis in the MAPS program, implying the utility of allowing backward passes. MARS first builds a model with its forward moves, and at the end of that process, a backward deletion procedure is applied. The D/S/A algorithm always is attempting to make backward and substitution selection moves throughout its search, thereby eliminating the luggage of undesirables. They conclude that polynomials are a reasonable choice for basis functions for reasons including its known approximation capabilities, interpretability, and

a model dimension which tends to be smaller than the sample size (Cox, 1988; Friedman, 1991, pg. 67-82). Polynomials served as a practical way to represent one version of the D/S/A algorithm. However, it is of interest to implement the D/S/A algorithm using spline basis functions and see how it behaves.

The D/S/A algorithm is currently implemented in C with subroutines from the NAG libraries, and it will be made into an R function. An overall description of the estimation methodology with examples of the D/S/A algorithms used here and in the context of histogram regression is available in (Dudoit et al., 2003).

References

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. The Wadsworth Statistics/Probability series. Wadsworth International Group, 1984.

R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, 2:65–73, 1998.

D. D. Cox. Approximation of least squares regression on nested subspaces. *The Annals of Statistics*, 16(2):713–732, 1988.

S. Dudoit and M. J. van der Laan. Asymptotics of cross-validated risk estimation in model selection and performance assessment. Technical Report 126, Division of Biostatistics, University of California, Berkeley, Feb. 2003. URL `www.bepress.com/ucbbiostat/paper126/`.

S. Dudoit, M. J. van der Laan, S. Keleş, A. M. Molinaro, S. E. Sinisi, and S. L. Teng. Loss-based estimation with cross-validation: Applications to microarray data analysis and motif finding. Technical Report 137, Division of Biostatistics, University of California, Berkeley, Dec. 2003. URL `www.bepress.com/ucbbiostat/paper137/`.

B. Durbin and S. Dudoit. A Deletion/Substitution/Addition algorithm for optimization of neural network architecture. Technical report, Division of Biostatistics, UC Berkeley, 2004. (In preparation).

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2), 2004.

J. H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–141, 1991. Discussion by A. R. Barron and X. Xiao.

S. Keleş, M. J. van der Laan, and M. B. Eisen. Identification of regulatory elements using a feature selection method. *Bioinformatics*, 18:1167–1175, 2002.

A. M. Molinaro and M. J. van der Laan. A Deletion/Substitution/Addition algorithm for partitioning the covariate space in prediction. Technical report, Division of Biostatistics, UC Berkeley, 2004. (In preparation).

I. Ruczinski, C. Kooperberg, and M. LeBlanc. Logic regression. *Journal of Computational and Graphical Statistics*, 12(3):475–511, 2003. URL `www.biostat.jhsph.edu/~iruczins/publications/publications.html`.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

S. E. Sinisi and M. J. van der Laan. Loss-based cross-validated Deletion/Substitution/Addition algorithms in estimation. Technical Report 143, Division of Biostatistics, University of California, Berkeley, March 2004. URL `www.bepress.com/ucbbiostat/paper143/`.

C. J. Stone, M. Hansen, C. Kooperberg, and Y. K. Truong. Polynomial splines and their tensor products in extended linear modeling. *The Annals of Statistics*, 25(4), 1997.

S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nature Genet.*, 22:281–285, 1999.

M. J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Technical Report 130, Division of Biostatistics, University of California, Berkeley, Nov. 2003. URL `www.bepress.com/ucbbiostat/paper130/`.

M. J. van der Laan and J. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer Series in Statistics. Springer, 2003.

M. J. van der Laan, S. Dudoit, and S. Keleş. Asymptotic optimality of likelihood based cross-validation. Technical Report 125, Division of Biostatistics, University of California, Berkeley, Feb. 2003. URL `www.bepress.com/ucbbiostat/paper125/`.

M. J. van der Laan, S. Dudoit, and A. W. van der Vaart. The cross-validated adaptive epsilon-net estimator. Technical Report 142, Division of Biostatistics, University of California, Berkeley, February 2004. URL `www.bepress.com/ucbbiostat/paper142/`.

J. van Helden, B. André, and J. Collado-Vides. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of

oligonucleotide frequencies. *Journal of Molecular Biology*, 281:827–842, 1998.

T. G. Wolfsberg, A. E. Gabrielian, M. J. Campbell, R. J. Cho, J. L. Spouge, and D. Landsman. Candidate regulatory sequence elements for cell cycle-dependent transcription in *Saccharomyces cerevisiae*. *Genome Research*, 9:775–792, 1999.

Chapter 4

Collaborative Targeted Maximum Likelihood Estimation

4.1 *Collaborative Double Robust Targeted Penalized Maximum Likelihood Estimation*

The following article appears as it was originally published on the University of California, Berkeley Division of Biostatistics Working Paper Series website in 2009, http://www.bepress.com/ucbbiostat/paper246/.

Collaborative Double Robust Targeted Penalized Maximum Likelihood Estimation

Mark J. van der Laan and Susan Gruber

Division of Biostatistics, University of California, Berkeley

Abstract

A new class of collaborative double robust targeted maximum likelihood estimators (C-DR-TMLE) targeting a particular parameter in a semiparametric model is proposed, building on the targeted maximum likelihood methodology of van der Laan and Rubin (2006). Targeted maximum likelihood estimation applies a targeted fluctuation function to a first stage (overall) density estimator and estimates the amount of fluctuation with parametric maximum likelihood estimation, treating the first stage density estimator as an offset. The optimal targeted fluctuation function typically depends on an unknown nuisance parameter.

In this article a fundamental further advance is achieved by generating a sequence of targeted maximum likelihood estimators with iincreasing likelihood indexed by increasingly nonparametric nuisance parameter estimators. Likelihood based cross-validation is used to select the nuisance parameter estimator for which the targeted maximum step yields the maximally effective bias reduction w.r.t. the target parameter.

A newly introduced collaborative double robustness of the efficient score equations solved by these targeted maximum likelihood estimators is shown to be superior to the current definition of double robustness in the estimating equation literature (e.g., Robins and Rotnitzky (2001) Robins et al. (2000), Robins (2000a), van der Laan and Robins (2003)), both in theory and in practice. As a consequence of this collaborative double robustness and maximum likelihood as the principal driving force, the resulting C-DR-TMLE is a more robust and optimal estimator of any pathwise differentiable parameter in any

semi-parametric model than the current state of the art in double robust estimation.

In addition, a general strategy of penalizing the log-likelihood so that the selection among different candidate targeted maximum likelihood estimators becomes more targeted towards the parameter of interest is introduced as well, which is able to avoid breakdowns of the estimation procedure for borderline identifiable target parameters. This results in a class of collaborative double robust targeted *penalized* maximum likelihood estimators (C-DR-TPMLE).

The method is illustrated in the context of estimation of causal effects in marginal structural models. In addition, simulations for nonparametric causal effect estimation illustrate the gain in practical performance of the collaborative double robust targeted maximum likelihood machine learning algorithms relative to current competitors such as the double robust estimating equation methodology that relies on an external non-collaborative estimator of the nuisance parameter. We also provide comparisons with ad hoc popular estimation procedures such as propensity score matching and inverse probability of treatment weighting. We also apply a particular C-DR-TPMLE implementation to assess the effects of mutations in the HIV virus on drug resistance.

This research provides a template for targeted efficient and robust machine learning of a particular target feature of the probability distribution of the data within large (infinite dimensional) semi-parametric models, while still providing statistical inference in terms of confidence intervals and p-values.

1 Introduction.

Researchers are beginning to acknowledge that questions about our infinite dimensional, semi-parametric world are not well-addressed by parametric models. More sophisticated tools are needed to wrest meaning from data. We can and should develop and utilize methods specifically designed to estimate a relatively small-dimensional precisely specified parameter within such a semiparametric model that is identifiable from the data. The ideal method would be entirely a priori specified, have desirable statistical properties, avoid reliance on ad hoc or arbitrary specifications, and be computationally feasible.

For example, suppose one observes a sample of independent and identically distributed observations from a particular data generating distribution in a semi-parametric model, and that one is concerned with estimation of a particular pathwise differentiable parameter of the data generating distribution. Due to the curse of dimensionality implied by the infinite dimension of semi-parametric models, standard (nonparametric) maximum likelihood often breaks down due to overfitting, and regularized sieve-based maximum likelihood estimation results in overly biased plug-in estimators of the parameter of interest.

The latter is due to the fact that such likelihood based estimators are aiming to estimate the density of the distribution of the data itself and thereby seek and achieve a bias-variance trade-off that is optimal for that whole density. Since the variance of an optimally smoothed density estimator is typically much larger than the variance of a smooth (pathwise-differentiable) parameter of the density estimator, the substitution estimators are often too biased relative to their variance. That is, substitution estimators based on density estimators involving optimal (e.g., likelihood-based) bias-variance trade-off (for the whole density) are not targeted towards the parameter of interest.

Motivated by this problem with the bias-variance trade-off of maximum likelihood estimation in semiparametric models, while still wanting to preserve the log-likelihood as the principle criterion in estimation, in van der Laan and Rubin (2006) we introduced and developed a targeted maximum likelihood estimator of the parameter of interest obtained by substitution of a targeted maximum likelihood estimator of the distribution of the data into the parameter mapping that maps the distribution of the data into the wished target parameter: i.e., it is still a plug-in estimator, but the density estimator is now targeted towards the parameter of interest.

The targeted maximum likelihood estimator of the distribution of the data is obtained by fluctuating an initial estimator of the data generating distribution with a parametric fluctuation model whose score at the initial estimator (i.e. at zero fluctuation) equals the efficient influence curve of the parameter of interest, and estimating the fluctuation parameter with maximum likelihood estimation, treating the initial estimator as fixed. The fluctuation model choice typically depends on an unknown nuisance parameter, which thus needs to be estimated as well. Iteration of this targeted maximum likelihood modification step results in a so called k-th step targeted maximum likelihood estimator, and its limit in k solves the actual efficient in-

fluence curve equation. The latter estimator we called the targeted maximum likelihood estimator, which also results in a corresponding plug-in targeted maximum likelihood estimator of the parameter of interest.

We refer to the log-likelihood of the targeted maximum likelihood estimator as the targeted log-likelihood, which provides a new loss function which can be used to evaluate candidate targeted maximum likelihood estimators, by applying cross-validation to the targeted log-likelihood loss function: see van der Laan and Rubin (2006) for a detailed exposition on the targeted log-likelihood loss function.

One can consider a targeted maximum likelihood estimator as a two stage estimator in which the first stage estimator is the initial (typically, non-targeted) estimator, and the second stage represents the updating of the initial estimator involving one or more iterative targeted maximum likelihood estimations along the fluctuation model, including the estimation of the nuisance parameter the fluctuation model depends upon. This targeted maximum likelihood step corresponds with fitting the parameter of interest and thereby results in considerable bias reduction (while increasing the likelihood), *if the nuisance parameter identifying this optimal fluctuation model is correctly specified.*

In van der Laan and Rubin (2006) we prove that targeted maximum likelihood estimators enjoy all the good properties of maximum likelihood estimators, but, in addition, they satisfy the double robust property of estimators based on solving optimal estimating equations in (e.g.) censored data models in which the censoring mechanism satisfies the coarsening at random assumption: that is, if either the first stage initial estimator or the nuisance parameter estimator (e.g., censoring mechanism) as required in the targeted maximum likelihood steps are consistent, then the resulting plug in estimator of the parameter of interest is consistent. In addition, targeted maximum likelihood estimators are locally efficient in the sense that they are asymptotically efficient (in the semiparametric model) if both the initial estimator as well as the nuisance parameter are consistently estimated, assuming the usual regularity conditions guaranteeing the convergence to a normal limit distribution. In fact, targeted maximum likelihood estimators are naturally super efficient if the initial estimator is consistent according to a (e.g.) parametric model and the nuisance parameter estimator is inconsistent by being based on a too small model (e.g., by failing to adjust for certain confounders): In such situations, the targeted maximum likelihood estimators essentially behaves as a parametric maximum likelihood estima-

tor according to a correctly specified parametric model (representing the augmentation of the original correctly specified parametric model with the targeted fluctuation function).

An outstanding open problem that obstructs the robust practical application of double robust estimators, including the targeted maximum likelihood estimators (in particular, in nonparametric censored data or causal inference models) is the selection of a sensible model or estimator of the nuisance parameter needed to evaluate the fluctuation model: this is especially the case when the efficient influence curve estimating equation involves inverse probability of censoring or treatment weighting, due to the enormous sensitivity of the estimator of the parameter of interest to the estimator of the censoring or treatment mechanism.

In this article we introduce an appealing new strategy for nuisance parameter estimator selection for targeted maximum likelihood estimators that addresses this challenge by taking the log-likelihood of the targeted maximum likelihood estimator indexed by the nuisance parameter estimator as the principal selection criterion. Our solution to nuisance parameter estimator selection is therefore a direct consequence of the introduction of targeted maximum likelihood estimation. Since even the nuisance parameter estimators as needed in the targeted maximum likelihood step are now based on the log-likelihood of the resulting targeted maximum likelihood estimator, the resulting estimator of the target parameter is completely likelihood based.

For the sake of illustration, while still covering a very wide range of statistical estimation problems, in this article we will focus on likelihoods that factor into a relevant factor and nuisance factor, as in censored data models satisfying the coarsening at random assumption (CAR). Even though a maximum likelihood estimator will only be concerned with the relevant factor of the likelihood identified by the full data distribution in a CAR censored data model, the targeting step in the targeted maximum likelihood estimators will depend on an estimator of the censoring mechanism. Our proposal is to select among candidate estimators or models of the censoring mechanism based on the cross-validated or empirical log-likelihood of the corresponding targeted maximum likelihood estimator implied by the initial estimator and the estimator of the censoring mechanism. In particular, we propose specific greedy algorithms for generating an estimator of the censoring mechanism that aims to maximize this targeted log-likelihood criterion among lots of candidate censoring mechanism estimators (indexed by different models).

We note that this estimator of the censoring mechanism targets a censoring mechanism that depends on the limit of the initial estimator of the relevant factor of the density of the data.

In addition, we propose to iterate this procedure resulting in the following explicit template for construction of our proposed collaborative targeted maximum likelihood estimators: 1) start with a first stage initial estimator of the relevant factor of the likelihood, 2) generate an estimator of the censoring mechanism based on an algorithm that maximizes, over candidate estimators of the censoring mechanism, the log-likelihood of the corresponding candidate targeted maximum likelihood estimators of the relevant factor, 3) select the resulting targeted maximum likelihood estimator at this particular selected estimator of the censoring mechanism, resulting in an update of the initial estimator, 4) iterate steps 1-3 (by using the update of 3) as initial estimator in 1)) to generate a sequence of targeted maximum likelihood estimators at increasingly nonparametric censoring mechanism estimators by maximizing the targeted log-likelihood as in 2) either over augmentations of the previously obtained fit of the censoring mechanism or over all candidate estimators that are more nonparametric than the previous one, and 5) use the cross-validated log-likelihood (or a penalized version as proposed here) to select among these candidate targeted maximum likelihood estimators indexed by the different censoring mechanism estimators (i.e., number of iterations), and possibly indexed by different initial estimators. Natural variations of this template can be included.

1.1 Organization of article

The complete description of the two stage collaborative targeted maximum likelihood methodology is presented in Section 2.

In Section 3 we present the statistical property of our proposed selector of the nuisance parameter estimator for the targeted maximum likelihood step by referring to the previously established oracle property of the likelihood based cross-validation selector among candidate density estimators. This exposition shows that our proposed methodology for building the second stage of the targeted maximum likelihood estimator is superior to alternate methods that estimate the nuisance parameter externally based on the likelihood for the nuisance parameter, as in the current literature.

The new method for estimation of the censoring mechanism (i.e., nuisance parameter) targets a true censoring mechanism that depends on the limit or

the rate of convergence of the initial estimator of the relevant factor, and thereby differs from the one implied by the log-likelihood of the censoring mechanism. As a consequence, this calls into question whether this type of targeted maximum likelihood estimator is still double robust. More general, what are the asymptotic properties of this new class of targeted maximum likelihood estimators, in particular, in relation to the regular targeted maximum likelihood estimator using an externally obtained nuisance parameter estimator?

Inspired by this, in Section 4 we establish a new double robustness property of the efficient influence curve (and estimating functions) in CAR censored data models that we name collaborative double robustness. This collaborative double robustness property provides additional confirmation (beyond the oracle property of the likelihood based cross-validation selector referred to in Section 3) of the validity of our approach for nuisance parameter estimation and represents a new approach to collaborative double robust targeted maximum likelihood estimation (C-DR-TMLE) with remarkable attractive properties.

The new collaborative double robustness is superior to the classical double robustness by providing additional important robustness: the censoring mechanism only needs to condition on confounders/covariates that have not been fully explained by the estimator of the relevant full-data distribution factor of the likelihood. We argue and show through simulations that, if the initial estimator is consistent, then this C-DR-TMLE is often more efficient than the locally efficient DR-TMLE: that is, the C-DR-TMLE is a so called super-efficient estimator that achieves an asymptotic variance that can be smaller than the variance of the efficient influence curve. In fact, our finite sample simulations show a remarkable superiority of the C-DR-TMLE relative to an efficient estimator like the DR-TMLE for estimation of a causal effect in the presence of confounding.

This implies that the C-DR-TMLE may be an irregular estimator at certain data generating distributions, since the best estimator among all regular estimators has asymptotic variance equal the variance of the efficient influence curve. A root-n consistent regular estimator is an estimator whose limit distribution stays the same under $\epsilon/\sqrt{n}$-fluctuations in the class of parametric fluctuations $\{P(\epsilon) : \epsilon\}$ of data generating distribution P, as used to define the path-wise derivative of the target parameter and its canonical gradient/efficient influence curve at P, for each P in the model. Indeed, we can argue that fluctuations in the true full data factor of the likelihood

of magnitude $\epsilon/\sqrt{n}$ can result in a different selection of estimator of the censoring mechanism and thereby results in a different limit distribution.

In Section 5 we show that under the fixed true data generating distribution the estimator can be expected to be asymptotically normally distributed and we provide tools for statistical inference based on its influence curve. Our simulations establish, in spite of the high sensitivity to the data of our estimator, reasonable coverage of the resulting confidence intervals.

From an asymptotic perspective the log-likelihood of the targeted maximum likelihood estimator is an excellent criterion to select among different targeted maximum likelihood estimators, possibly indexed by different candidate estimators of the nuisance parameters (e.g., censoring mechanism). In particular, it results in the new form of collaborative double robustness of the resulting collaborative targeted maximum likelihood estimators. However, we argue and show that in various applications in which the target parameter is borderline identifiable, for the purpose of nuisance parameter estimator selection, the targeted log-likelihood of a density estimator may not be sensitive enough towards the mean squared error of the substitution estimator (corresponding with the density estimator) of the parameter of interest.

Therefore in Section 6 we propose a targeted penalized log-likelihood based criterion that can instead be used to select among targeted maximum likelihood estimators indexed by an index δ, for example, representing different choices for the nuisance parameter estimator. Our penalty for the log-likelihood at a density estimator concerns an estimator of a mean squared error of the substitution estimator of the target parameter. In particular, we propose to estimate this mean squared error with the covariance matrix of the efficient influence curve of the parameter of interest at a candidate (targeted maximum likelihood) density estimator, and a bias estimate of the substitution estimator of the target parameter relative to its asymptotic limit. One can also use the bootstrap to estimate this mean squared error term, if computer resources allow. The penalty is scaled appropriately and so that the penalized log-likelihood criterion is asymptotically dominated by the log-likelihood criterion.

The proposed criterion for selection among two stage δ-specific targeted maximum likelihood estimators equals the sum of the cross-validated log-likelihood of the δ-specific targeted maximum likelihood estimator and a mean squared error term (relative to its δ-limit). The mean squared error term is split up as a sum of an appropriately scaled function of the cross-

validated or empirical estimate of the variance of the efficient influence curve at the δ-specific targeted maximum likelihood estimator, and the square of a cross-validated estimate of the bias in the δ-specific targeted maximum likelihood estimator relative to its limit for fixed δ. We also consider splitting up the cross-validated log-likelihood into a sum of the cross-validated log-likelihood of the δ-specific initial maximum likelihood estimator and the gain in empirical or cross-validated log-likelihood due to the targeted maximum likelihood step for the δ-specific targeted maximum likelihood estimator. In this separation the first term concerns the performance of the initial estimator while the second term concerns the performance of the targeted maximum likelihood estimation applied to the initial estimator.

We provide rationales for each of the terms in the proposed penalized targeted log-likelihood criterion, which demonstrate that it is able to deal with a variety of challenges that come with selection among density estimators of the data generating distribution w.r.t. the parameter of interest of the true density, without affecting the likelihood as principle criterion.

In Section 7 we consider estimation of a causal effect in a marginal structural model to illustrate the importance of this finite sample parameter specific penalty term for the log-likelihood and define the collaborative double robust targeted penalized maximum likelihood estimator of the unknown parameters of the marginal structural model.

In Section 8 we consider estimation of a causal effect of a binary treatment in a nonparametric model, and present a particular implementation of C-TMLE used in our simulations and data analysis. We illustrate the superior performance of our C-DR-TMLE in this latter estimation problem in comparison with various competitors representing current practice.

In Section 9 we carry out a data analysis and again observe excellent finite sample performance.A discussion is presented in Section 10. Section 11 provides extensions to the core methodology.

2 Collaborative double robust targeted maximum likelihood estimators.

We will describe our newly proposed targeted maximum likelihood estimators in the context of censored data models, but the generalization to general semi-parametric models is immediate.

Let $O = \Phi(C, X)$ be a censored data structure on a full data random variable X, where C denotes the censoring variable. We assume coarsening at random so that the observed data structure $O \sim P_0$ has a probability distribution whose density w.r.t an appropriate dominating measure factors as $dP_0(O) = Q_0(O)g_0(O \mid X)$, where Q_0 is the part of the distribution of X that is identifiable, and g_0 denotes the conditional probability distribution of O, given X, which we often refer to as the censoring mechanism. By CAR, we have $g_0(O \mid X) = h(O)$ for some measurable function h. If C is observed itself, then g_0 denotes the conditional distribution of C, given X.

A semiparametric model $\mathcal{M}$ for the probability distribution P_0 of the observed data structure O is implied by a model $\mathcal{Q}$ for the full-data distribution factor Q_0, and a model $\mathcal{G}$ for the censoring mechanism g_0. Let $O_1, \ldots, O_n$ be n independent and identically distributed (i.i.d.) observations of the experimental unit O with probability distribution $P_0 \in \mathcal{M}$.

Let $\Psi : \mathcal{M} \to \mathbb{R}^d$ be a d-dimensional parameter that is path-wise differentiable at each $P \in \mathcal{M}$ (w.r.t. a class of finite dimensional paths through P) with efficient influence curve $D^*(P)$. For the sake of illustration, it is assumed that $\Psi(P_{Q,g}) = \Psi^F(Q)$ for some Ψ^F: i..e, the parameter of interest is a parameter of the full data distribution of X. The efficient influence curve $D^*(P)$ at P with $dP = Qg$ will also be denoted with $D^*(Q, g)$.

Let P_n be the empirical probability distribution of $O_1, \ldots, O_n$ which puts mass $1/n$ on each of the n observations.

The Targeted Maximum Likelihood estimator indexed by initial (Q, g): Given any $P \in \mathcal{M}$ with $dP = Qg$, let $\{P(\epsilon) : \epsilon\} \subset \mathcal{M}$ be a submodel with finite dimensional parameter ϵ, dominated by P, through P at $\epsilon = 0$, and whose scores at $\epsilon = 0$ span a finite dimensional space within $L_0^2(P)$ that includes the efficient influence curve $D^*(P) = D^*(Q, g)$. Because our parameter of interest is a parameter of Q_0 and the factorization $dP_0 = Q_0 g_0$, it follows that such a fluctuation model can be chosen to only fluctuate Q with a submodel $Q_g(\epsilon) \subset \mathcal{Q}$, where this fluctuation model will be indexed by g. Let $dP(\epsilon) = Q_g(\epsilon)g$ be such a fluctuation model with fluctuation parameter ϵ.

At a given (Q, g), one can now define a k-th step targeted maximum likelihood version $Q_g^k(P_n)$ of Q_0 as follows. Firstly, let $Q_g^1(P_n) = Q_g(\epsilon_n^1)$, where

$$\epsilon_n^1 = \arg\max_{\epsilon} P_n \log Q_g(\epsilon).$$

Here we use the notation $Pf = \int f(o)dP(o)$. In general, $Q_{gn}^k = Q_g^k(P_n) =$

$Q_g^{k-1}(P_n)(\epsilon_n^k)$, where

$$\epsilon_n^k = \arg\max_{\epsilon} P_n \log Q_g^{k-1}(P_n)(\epsilon).$$

One iterates this updating till ϵ_n^k equals zero within a user supplied precision. The final update is refered to as the (iterative) targeted maximum likelihood estimator $Q_{gn}^* = Q_g^*(P_n)$, indexed by the initial starting point (Q, g).

The Targeted Maximum Likelihood estimator indexed by initial estimator and estimator of nuisance parameter: The above procedure, applied to an initial estimator Q_n^0, and an estimator g_n of g_0, defines the k-th step targeted maximum likelihood estimator and its limit in k, Q_n^*, as introduced and analyzed in van der Laan and Rubin (2006).

Estimation of the censoring mechanism for TMLE-update of initial estimator: However, an important consideration and open problem addressed in this article is how to select the estimator g_n that defines the targeted maximum likelihood update of the first stage estimator Q_n^0. For example, what measured covariates should one adjust for in this censoring mechanism model?

In this article we propose the following (class of) algorithms that provide a likelihood driven method for generating such an estimator g_n of censoring mechanism g_0, but also provides a further augmentation of the above targeted maximum likelihood approach by building a sequence of second stage targeted maximum likelihood nested bias reductions to choose from, and using the cross-validated log-likelihood criterion or a penalized cross-validated log-likelihood (more targeted towards the target parameter) to choose among them.

2.1 Collaborative Targeted MLEs.

We present the following template providing a new class of so called collaborative targeted maximum likelihood estimators.

Candidate estimators of censoring mechanism: For each δ in an index set, let $g_{n\delta}$ be a candidate estimator of g_0. Let $d(\delta)$ denote a measure of how data adaptive $g_{n\delta}$ is, and for a maximal value $d(\delta)$ or for $d(\delta)$ approximating a maximum value we have that $g_{n\delta}$ is actually a consistent estimator of g_0.

For example, let $\mathcal{G}_\delta \subset \mathcal{G}$ be a submodel indexed by an index δ ranging over an index set, and let $d(\delta)$ denote a dimension of $\mathcal{G}_\delta$ so that $\mathcal{G}_\delta$ approximates $\mathcal{G}$ for $d(\delta)$ converging to infinity. Thus $\{\mathcal{G}_\delta : \delta\}$ denotes a sieve for the model $\mathcal{G}$ for g_0. Now, $g_{n\delta}$ could be defined as (a possibly regularized) maximum likelihood estimator of g_0 according to this model $\mathcal{G}_\delta$.

A particular type of candidate estimator $g_{n\delta}$ is indexed by a data adaptive ordering of a set of covariates/confounders, and a maximum likelihood based machine learning algorithm, such as the super learning algorithm of van der Laan et al. (2007) which obtains an optimal weighted combination of a user supplied set of candidate machine learning algorithms, for estimating the censoring mechanism based on the set consisting of the first $k = \delta$ covariates.

For example, the data adaptive ordering could be based on a forward greedy algorithm aiming to find the main terms that need to be included in the censoring mechanism to provide effective targeted maximum likelihood bias reduction for the target paramater, starting with the initial estimator. In this manner, the first covariates in the ordering are likely the most important confounders to adjust for in collaboration with the initial estimator.

Another example is to order the covariates by the correlation with the censoring variable so that potentially harmful covariates are pushed towards the end of the ordered sequence of covariates.

In the above few examples, the collection of δ-values is finite and not that large. Alternatively, δ identifies a set of basis functions used in a linear model for g_0 and $g_{n\delta}$ is the corresponding maximum likelihood estimator. In this case, the collection of δ values is typically extremely large, so that heuristic greedy algorithms will be needed in the next step.

Censoring mechanism estimator by maximizing targeted log-likelihood: Consider an algorithm that takes as input a starting choice δ^*, and searches among a specified set of candidate estimators $g_{n\delta}$ with $d(\delta) > d(\delta^*)$ with the goal of maximizing the targeted log-likelihood criterion.

$$\delta \to P_n \log Q^*_{g_{n\delta}}(P_n). \tag{1}$$

Recall that $Q_g^*(P_n)$ denotes the iterative targeted maximum likelihood estimator that uses the optimal fluctuation model identified by censoring mechanism g. Note that this algorithm will select an estimator of the censoring mechanism that is more nonparametric than its starting estimator $g_{n\delta^*}$, and that its corresponding targeted maximum likelihood estimator has a larger empirical targeted log-likelihood than the targeted maximum likelihood estimator indexed by $g_{n\delta^*}$.

First step collaborative targeted maximum likelihood estimator: This provides us now with the ingredients to define our first step collaborative targeted maximum likelihood estimator. We start the above algorithm with a δ_{start} with $d(\delta_{start}) = 0$, and suppose that it ends up with a choice δ_n^1, and thereby an estimator $g_n^1 = g_{n\delta_n^1}$. This defines now a first-step targeted maximum likelihood estimator $Q_n^{*1} = Q_{g_n^1 n}^*$, which satisfies that its empirical log-likelihood is larger than the empirical log-likelihood of the initial estimator: $P_n \log Q_n^{*1} > P_n \log Q_n^0$.

This defines a mapping that takes as input $Q_{start} = Q_n^0$ and $g_n^0 = g_{\delta_{start}}$ (say the intercept model), and maps it into a targeted estimator Q_n^{*1} and corresponding censoring mechanism estimator g_n^1, where this mapping involves applying an algorithm searching and maximizing over candidate (e.g., maximum likelihood estimators) $\{g_{n\delta} : \delta\}$, w.r.t. the log-likelihood of the corresponding targeted maximum likelihood estimator of Q_0 starting at Q_n^0, and subsequently selecting the resulting targeted maximum likelihood estimator at the selected estimator of the censoring mechanism.

We will refer to this estimator Q_n^{*1} or the pair (Q_n^{*1}, g_n^1) as a first step collaborative targeted maximum likelihood estimator, using the term "collaborative" to indicate that the estimator Q_n^{*1} involves a collaboration with the initial estimator Q_n^0 when determining the estimator g_n^1 of the censoring mechanism g_0.

Iteration. The k-th step collaborative targeted maximum likelihood estimator (C-TMLE): We can now iterate this process of mapping an initial estimator Q_n^0 and g_n^0 into a targeted estimator Q_n^{*1} and corresponding censoring mechanism estimator g_n^1: set $Q_{start} = Q_n^{*1}$, $\delta_{start} = d(g_n^1)$, compute Q_n^{*2}, g_n^2, set $Q_{start} = Q_n^{*2}$, and so on, giving us a sequence of estimators (Q_n^{*k}, g_n^k), $k = 1, \ldots$, where g_n^k denotes the censoring estimators selected in k-th step of this iterative algorithm.

We refer to the estimator Q_n^{*k} and its pair (Q_n^{*k}, g_n^k) as the k-th step collaborative targeted maximum likelihood estimator. Beyond these candidate estimators indexed by k, one can also define Q_n^* as the limit in k of Q_n^{*k}, which will correspond with iteration till the point that the most nonparametric estimator $g_n = g_{nk}$ has been selected.

Possible refinement/alternative of candidate C-TMLEs: We can define a set of values $d_1 < \ldots < d_K$, and determine, for each $k = 1, \ldots, K$, the above estimator Q_n^* but restricting our set of candidate estimators of the censoring mechanism to $\{g_{n\delta} : d(\delta) \leq d_k\}$, all censoring mechanism estimators with complexity measure less than or equal to d_k. This now generates K candidate C-TMLE's Q_n^{*k} and corresponding (the in final step selected censoring mechanism estimator) g_n^k, $k = 1, \ldots, K$.

Note that in this way, we may obtain a richer set of candidate estimators representing a larger set of possible nested bias reductions, resulting in a more refined bias-variance trade-off when using likelihood based cross-validation to select among these candidate estimators.

For example, when applied to the data set, Q_n^{*1} might only carry out one targeted maximum likelihood step selecting $g_{n\delta}$ with $d(\delta) = d_1$, while Q_n^{*2} might still only represent one targeted maximum likelihood step but now selecting $g_{n\delta}$ with $d(\delta) = d_2$, and, say, Q_n^{*3} will carry out two targeted maximum likelihood steps, and so on.

Other variations of such proposed candidate C-TMLE's may be appreciated as well. For example, if one uses a greedy forward (bottom up) algorithm when aiming to maximize the targeted log-likelihood criterion (1) (over candidate censoring mechanism estimators), generating a sequence of censoring mechanism estimators of increasing size $1, 2, 3, \ldots$, moving towards the next targeted maximum likelihood iteration when the local maximum has been achieved, then each size indexes a corresponding candidate C-TMLE in which the last targeted maximum likelihood iteration uses the censoring mechanism estimator of that size. Such a sequence of candidate C-TMLEs is presented and implemented in our simulation section.

All candidate C-TMLEs solve an efficient influence curve equation: Since all candidate C-TMLE's are targeted maximum likelihood estimators using the last selected censoring mechanism estimator to carry

out the (possibly iterative) targeted maximum likelihood algorithm, we have that all of them solve the efficient influence curve equation:

$$0 = P_n D^*(Q_n^{*k}, g_n^k),\ k = 1, \ldots, K.$$

This is a fundamental property of our candidate collaborative targeted MLEs driving the targeted bias reduction w.r.t. the target parameter of interest.

Cross-validation to select number of iterations k in k-th step C-TMLE: Given this sequence of candidate collaborative targeted maximum likelihood estimators $P_n \to (Q_n^{k*} =)\hat{Q}^{k*}(P_n)$ indexed by k, it remains to select k.

We select k based on the cross-validated log-likelihood:

$$k_n = \underset{k}{\text{argmax}}\ E_{B_n} P^1_{n,B_n} \log \hat{Q}^{k*}(P^0_{n,B_n}),$$

where the random vector $B_n \in \{0,1\}^n$ denotes a cross-validation scheme such as V-fold cross-validation, and P^0_{n,B_n}, P^1_{n,B_n} are the empirical probability distributions of the training sample $\{i : B_n(i) = 0\}$ and validation sample $\{i : B_n(i) = 1\}$, respectively, as identified by the split vector B_n.

The Collaborative (Double Robust) Targeted Maximum Likelihood Estimator: The corresponding targeted maximum likelihood estimator of $\psi_0 = \Psi^F(Q_0)$ is given by

$$\Psi(Q_n^{k_n*}) = \Psi(\hat{Q}^{k_n*}(P_n)).$$

We refer to this estimator as the collaborative (double robust) targeted maximum likelihood estimator (C-DR-TMLE or C-TMLE).

Additional fine tuning of the second stage of the C-DR-TMLE: If the above algorithm for maximizing the targeted log-likelihood over candidate censoring mechanism estimators is indexed by a choice δ_2, then our candidate C-DR-TMLE are indexed by a choice δ_2 and k. In this case we would select both k and δ_2 as we selected k above, based on the cross-validated log-likelihood:

$$(\delta_{2n}, k_n) = \underset{\delta_2, k}{\text{argmax}}\ E_{B_n} P^1_{n,B_n} \log \hat{Q}^{k\delta_2*}(P^0_{n,B_n}).$$

Possible joint fine tuning of first stage and second stage in the C-DR-TMLE: If the first stage initial estimator is indexed by different choices δ_1, then one selects both the first stage estimator choice δ_1 as well as the second stage choices (k, δ_2) based on the cross-validated log-likelihood of the actual targeted maximum likelihood estimator:

$$(\delta_{1n}, \delta_{2n}, k_n) = \underset{\delta_1,\delta_2,k}{\operatorname{argmax}} E_{B_n} P^1_{n,B_n} \log \hat{Q}^{k\delta_2*}_{\delta_1}(P^0_{n,B_n}).$$

Separation of First Stage and Second Stage fine tuning in the C-DR-TMLE: If the first stage initial estimator is indexed by different choices δ_1, then it is also possible to select δ_1 based on the cross-validated log-likelihood of such initial estimators $P_n \to \hat{Q}_{\delta_1}(P_n)$, so that in the above description Q^0_n is simply $\hat{Q}_{\delta_{1n}}(P_n)$ with δ_{1n} defined as

$$\delta_{1n} = \underset{\delta_1}{\operatorname{argmax}} E_{B_n} P^1_{n,B_n} \log \hat{Q}_{\delta_1}(P^0_{n,B_n}).$$

Penalized cross-validated log-likelihood: The above cross-validation selection can be replaced by cross-validation based on a more targeted loss function and or a cross-validated penalized log-likelihood, as we present in a later section. Such a criterion for a density estimator is selected to be more sensitive towards the behavior of the corresponding substitution estimator of the target parameter, without affecting its capability to select for overall fits. In a later section we propose a penalty based on an estimator of the mean squared error of the substitution estimator relative to its asymptotic limit. In particular, such a mean squared error could be estimated with the bootstrap, but we provide analytic formulas that can be used instead as well.

Super Learning to select among different C-DR-TMLE: For simplicity, let $\hat{Q}^{j*}(P_n), \hat{g}^j(P_n)$, $j = 1, \ldots, J$, denote all candidate C-TMLE $\hat{Q}^{j*}(P_n)$ with corresponding censoring mechanism estimator $\hat{g}^j(P_n)$, indexed by different second stage algorithms (e.g. indexed by number of iterations, choice of algorithm, etc). One could now define new candidates $\sum_j \alpha(j)\hat{Q}^{j*}(P_n)$ indexed by weight vectors α. Even though each targeted maximum likelihood estimator $\hat{Q}^{j*}(P_n)$ solves the efficient influence curve equation, $0 = P_n D^*(\hat{Q}^{j*}(P_n), \hat{g}^j(P_n))$, viewed as a requirement for bias reduction w.r.t. target parameter, this does not imply

that weighted combinations of targeted maximum likelihood estimators also solve an efficient influence curve equation. Therefore, for each combination we would still need to define an (iterative) targeted maximum likelihood step, at an appropriately selected censoring mechanism estimator g_n^α, that maps the combination estimator $\sum_j \alpha(j)\hat{Q}^{j*}(P_n)$ into a solution of the efficient influence curve equation

$$0 = P_n D^* \left(\sum_j \alpha(j)\hat{Q}^{j*}(P_n), g_n^\alpha \right).$$

One could now select α by maximizing the cross-validated log-likelihood of these α-specific targeted maximum likelihood estimators.

In this way, one can use super learning to find the optimized combination of candidate collaborative targeted maximum likelihood estimators indexed by different possible choices of initial estimator and or second stage targeted bias reduction.

C-TMLE with unordered candidate nuisance parameter estimators: The template outlined above relies upon knowing for any given candidate nuisance parameter estimator, representing the previously selected nuisance parameter estimator at the current C-TMLE step, a set of candidates that are more nonparametric than the given nuisance parameter estimator.

Suppose we are given a set of candidate nuisance parameter estimators that include heavily nonparametric estimators, but for which we do not know how to order them from least to most non-parametric. We start out with selecting one nuisance parameter estimator from this list of candidates resulting in the first step C-TMLE as above. In the second step C-TMLE we now consider nuisance parameter estimators that are combinations of the current nuisance parameter estimator with any of the candidates. It is assumed that the combinations are always more nonparametric than the candidate nuisance parameter estimators that are combined: for example, the combinations are convex combinations of the current nuisance parameter estimator with a candidate with the coefficients being fitted with maximum likelihood. As possible variations one might decide to refit the coefficients of the previous combinations, and, to exclude already selected nuisance parameter es-

timators from the list. In this way the above template for C-TMLE can be carried out.

2.2 The rationale of the collaborative-TMLE

For simplicity consider the case that we are given candidates $g_{n\delta}$ for a finite set of possible δ, and that one of the candidates actually converges to the true g_0. It follows that for k large enough g_n^k will converge to the true g_0.

A few principles drive the asymptotic properties of the C-TMLE. Firstly, a TMLE using g_0 in the second stage to carry out the targeted maximum likelihood steps will be consistent for ψ_0.

Secondly, the asymptotic limit of the log-likelihood, $P_0 \log Q_n^{*k}$, for $n \to \infty$, is increasing in k as long as it is possible to increase the asymptotic log-likelihood criterion: e.g, as long as there remain true confounders that have not yet been properly adjusted for in either the current Q_0-fit or the censoring mechanism fit.

As a consequence, a cross-validated log-likelihood criterion will select larger and larger k_n for $n \to \infty$ as long as it can result in a true log-likelihood increase for the Q_0-fit. Therefore for large enough sample size one either will start using an estimator g_n that converges to the true g_0, or the estimator $Q_n^{*k_n}$ already fits the targeted maximum likelihood direction implied by the g_0-path, and less nonparametric estimators g_n will be selected that converge to some $g_0(Q)$. In either case, we will have that the limit Q of the selected collaborative targeted maximum likelihood estimator $Q_n^* = Q_n^{*k_n}$ will satisfy $\Psi(Q) = \psi_0$. Formally, the limit $g_0(Q)$ of the selected g_n satisfies the wished property

$$P_0 D^*(Q, g_0(Q)) = 0 \text{ implies } \Psi(Q) = \psi_0.$$

We will discuss this collaborative double robustness property of the efficient influence curve D^* in detail in the next section.

To summarize, the idea is that one iteratively applies the targeted maximum likelihood updating algorithm at more and more nonparametric models fits of g_0, thereby generating a sequence of k-th step collaborative targeted maximum likelihood estimators Q_n^{k*} based on more and more nonparametric fits g_n^k as k increases. This template covers the case where the censoring mechanism is a vector of censoring variables, e.g. treatment and missingness and right censoring. For example, g can be the conditional distribution of action process $A()$ given the full data, where $A(t)$ measures both miss-

ingness and treatment actions at time t. In addition, by construction, the log-likelihood of Q_n^{k*} is increasing in k since Q_n^{k*} is a targeted maximum likelihood estimator applied to initial estimator Q_n^{k-1*} (with censoring mechanism fit g_n^{k-1}).

By using the cross-validated log-likelihood to select k one will not over-select k and thus will only select meaningful censoring mechanism estimators that increase the fit of Q_0 during the targeted maximum likelihood algorithm. At the same time, one will also not under-select k since the most nonparametric censoring mechanism estimator is included as a candidate that will only be ignored if its targeted maximum likelihood direction is already fitted well by Q_n^{*k} for smaller k.

Figure 1 illustrates the collaborate nature of the construction of the sequence increasingly data-adaptive nuisance parameter estimators, $\{g_n^1, \ldots g_n^K\}$. A plot of the density of a poor initial estimator of Q_0 applied to $n = 5000$ simulated observations and the corresponding sequence of censoring mechanism estimators is shown in the to half of the figure. When the initial fit is poor, the nuisance parameter estimator converges quickly to g_0, and the selected candidate estimator closely approximates g_0. Plots in the bottom half of the figure shows the behavior of the C-TMLE procedure when Q_n^0 is well-estimated. When the initial fit is good, the nuisance parameter estimator grows slowly towards g_0, and the selected candidate estimator captures only a portion of the true nuisance parameter.

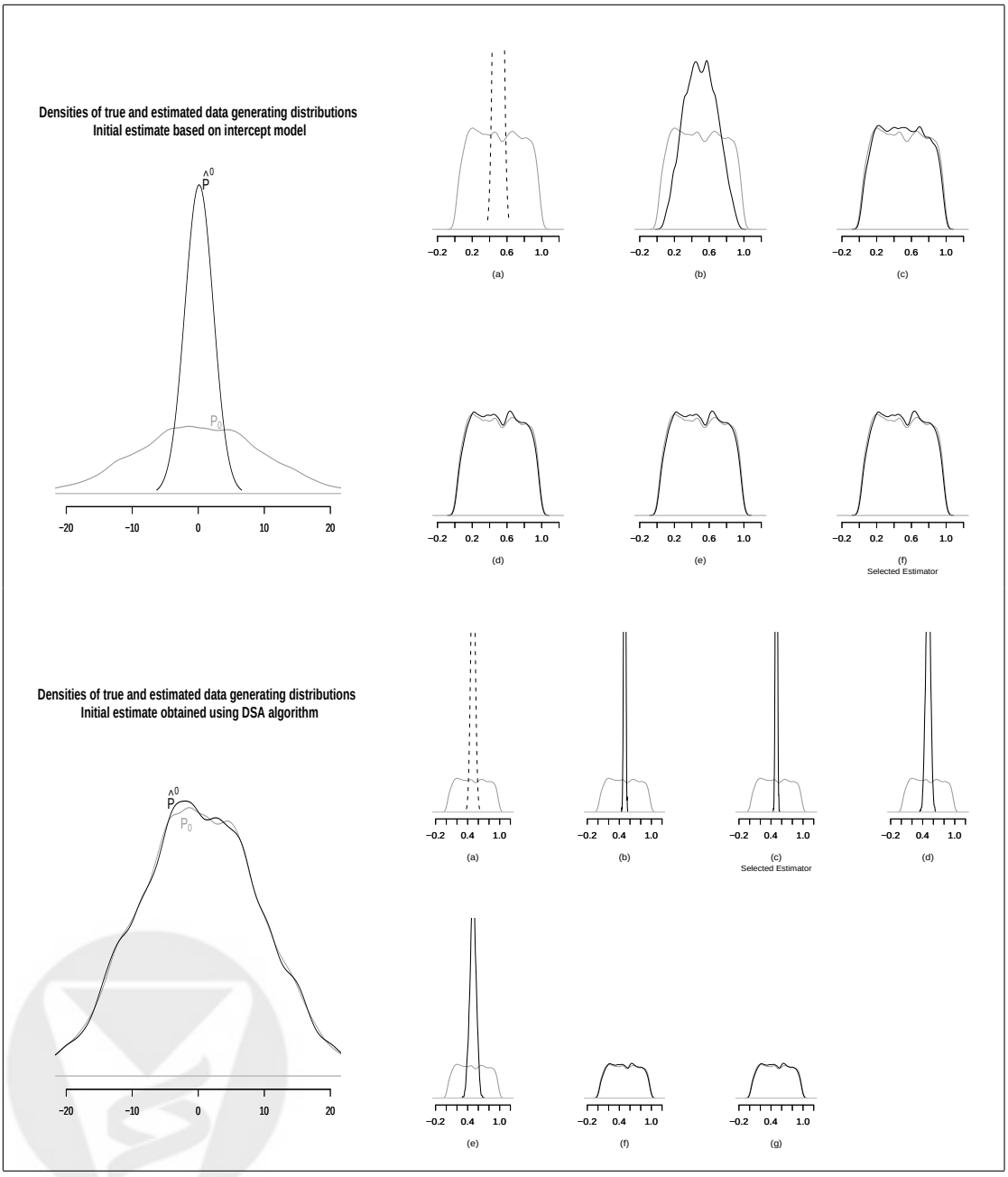


Figure 1: Construction of a sequence of nuisance parameter estimators based on a poor initial fit of the density (top) and a good initial fit for the density (bottom). True densities Q_0 and g_0 are shown in gray.

3 The superiority of the collaborative targeted MLE relative to a targeted MLE using an external estimator of censoring mechanism.

Our collaborative targeted MLE can be presented in words as follows. Firstly, we create candidate two-stage targeted maximum likelihood estimators of Q_0 that are defined as the iterative application of a targeted maximum likelihood step, starting with an initial first stage estimator, with an increasingly nonparametric choice of censoring mechanism estimator to estimate the optimal direction/fluctuation used in the targeted maximum likelihood step. This results in candidate k-th step targeted maximum likelihood estimators using more and more nonparametric estimators of the censoring mechanism to identify the optimal targeted direction, $k = 1, \ldots, K$, all starting with the same first stage estimator but differing in the number of iterations defining the second stage of the estimator.

At each targeted maximum likelihood step, the choice of censoring mechanism is based on maximizing the targeted log-likelihood gain and a set of candidates to choose from, where the set of candidates to choose from needs to be more nonparametric than the censoring mechanism estimator selected in the previous step. For example, we might a priori generate a set of candidate maximum likelihood estimators of the censoring mechanism based on a sequence of models for the censoring mechanism that approximates the true (large) model for the censoring mechanism (i.e., a so called sieve).

The collaborative targeted MLE is now defined by selecting k based on the cross-validated targeted log-likelihood for Q_0. This choice k_n of the cross-validation selector also identifies a choice of censoring mechanism estimator, namely the one used in the k_n-th targeted maximum likelihood algorithm. This choice of censoring mechanism estimator tells us how agressively our proposed collaborative targeted MLE pursues the bias reduction. This particular way of selecting the censoring mechanism estimator has strong implications for the practial and theoretical behavior of the collaborative targeted MLE of ψ_0 relative to a single step targeted MLE based on an external estimator g_n. These implications will be discussed in this section.

Firstly, we are concerned with understanding the statistical property of the selection procedure defining this choice of censoring mechanism estima-

tor. Under the assumption that the log-likelihood (i.e., log-Q) loss function is uniformly bounded among all candidate estimators of Q_0, and the number of candidates is polynomial in sample size, we can apply the theoretical results for the likelihood based cross-validation selector k_n of k as presented in van der Laan et al. (2004). These theoretical results teach us the cross-validation selector k_n will have the following statistical property: if none of the candidate two stage targeted maximum likelihood estimators of Q_0 achieve the parametric rate $1/\sqrt{n}$-rate of convergence, then the resulting estimator of Q_0 is asymptotically equivalent (not only in rate but also including the constant) with the oracle selector that for each data set selects the k that minimizes the Kullback-Leibler dissimilarity with Q_0, else, the resulting estimator will achieve (at minimal) the almost parametric rate of convergence $\log n/\sqrt{n}$. We are now ready to evaluate the implications of this theoretically established optimality property of the cross-validation selector for the resulting censoring mechanism estimator g_n, and thereby for the bias of the resulting estimator of ψ_0.

For simplicity, let's first only consider the comparison between the initial estimator and the first step collaborative targeted maximum likelihood estimator, using g_n^1 as estimator of the censoring mechanism in the targeted maximum likelihood algorithm. By iterating the argument our claims will follow. Firstly, consider the case that g_0 is known and the cross-validation selector only selects between the first step collaborative targeted maximum likelihood estimator using g_0, and the initial estimator. The cross-validation criterion will favor the first step estimator relative to the initial estimator Q_n if the first targeted direction was not fitted yet by the initial estimator Q_n, since it will bring the fit closer to Q_0. By the same argument, if the estimator g_n^1 converges faster to a fixed g_0^1 than Q_n converges to Q_0, and this direction identified by g_{01} was not fitted yet by Q_n, then the cross-validation criterion will favor the first step again. However, if the estimator g_n^1 converges slower to g_0^1 than the initial Q_n converges to Q_0, then the rate of convergence of the first step targeted maximum likelihood estimator to Q_0 will be worse than the initial estimator, so that the oracle selector and thereby the cross-validation selector will not favor the first step C-TMLE. By analogy, this argument shows that the cross-validation selector will generally favor higher step targeted maximum likelihood estimators until the corresponding estimator g_n^k converges at a slower rate to a fixed g_0^k than the $k-1$-step collaborative targeted maximum likelihood estimator converges to Q_0.

To summarize: The two stage k-th step collaborative targeted maximum likelihood estimators involves building a two stage estimator of Q_0. If the directions targeted in the second stage are estimated at a worse rate than the rate at which the initial estimator converges to Q_0, then the second stage fit is slowing down the rate of convergence of the two stage estimator. The oracle selector among all the two stage estimators, and thereby the cross-validation selector, will not allow this to happen by definition of the oracle selector. On the other hand, if the maximum likelihood directions targeted in the second stage are estimated at a precision going beyond the precision of the first stage estimator, then the oracle selector, and thereby the cross-validation selector, will favor such a second stage estimator.

In contrast, an externally estimated censoring mechanism can easily yield a second stage component having a worse convergence rate than the first stage, and thereby a resulting targeted MLE that is, in fact, a worse estimator than the initial estimator. This shows the enormously important power of selecting the censoring mechanism estimator with our proposed methodology based on the cross-validated log-likelihood for the resulting collaborative targeted maximum likelihood estimator of Q_0.

Note that this can result in the following scenarios. Firstly, suppose that the initial estimator Q_n is doing well and that the true g_0 is extremely hard to estimate. For example, g_0 might be a highly non-smooth function. Suppose also that there is an a priori set of candidate estimators $g_{n\delta}$ indexed by a choice δ, including less aggressive estimators that involve less adjustment but are converging fast. Then our proposed two stage collaborative targeted maximum likelihood estimator will still involve second stage targeted maximum likelihood algorithms at reasonable estimators $g_{n\delta}$, but it will avoid carrying out the noisy (relative to precision of initial estimator) targeted maximum likelihood algorithm corresponding with the nonparametric estimator g_n of g_0.

In particular, if the first stage estimator Q_n converges at a parametric rate to Q_0, then the cross-validation selector will select a censoring mechanism estimator g_n among the candidates that converges at a parametric or at most the almost parametric $\log(n)/\sqrt{n}$ rate. That is, in this case the selected censoring mechanism estimator g_n will not adjust for certain variables in order to keep the rate of convergence for the two stage estimator close to parametric. In the latter case, the resulting two stage collaborative targeted maximum likelihood estimator of ψ_0 would result in a super efficient estimator, since the estimator will be asymptotically normal at root-n rate

with smaller asymptotic variance than the variance of the efficient influence curve.

We also note that in the case that the initial estimator Q_n and the selected censoring mechanism estimator g_n are such that no parametric rate is achieved for ψ_n, then we can have a relative efficiency of infinity in favor of the collaborative targeted MLE relative to a regular double robust targeted MLE using an externally selected g_n that converges at a slower rate than Q_n. The reason is that the slower rate of g_n and thereby the second stage component of the estimator will now actually bring down the rate of convergence of the targeted maximum likelihood estimator ψ_n relative to the rate of convergence of the collaborative targeted MLE.

3.1 Formalizing the claimed property of the cross-validation selector of the censoring mechanism estimator in the collaborative targeted MLE.

Let Q_n^* be a targeted maximum likelihood estimator (playing the role of one of the candidate two stage collaborative targeted maximum likelihood estimators) using Q_n as an initial estimator. Let $d(Q, Q_0) = E_0 \log Q_0/Q$ be the Kullback-Leibler dissimilarity between a candidate Q and the true Q_0. The oracle selector will compare the performance measures $d(Q_n^*, Q_0)$ and $d(Q_n, Q_0)$ and prefer the one with the best performance, and as stated above, the cross-validation selector will follow the oracle selector closely. Therefore, we consider

$$\begin{aligned} \frac{d(Q_n^*, Q_0)}{d(Q_n, Q_0)} &= \frac{d(Q_n, Q_0) + d(Q_n^*, Q_0) - d(Q_n, Q_0)}{d(Q_n, Q_0)} \\ &= 1 - \frac{E_0 \log Q_n^*/Q_n}{E_0 \log Q_n/Q_0}. \end{aligned}$$

If this relative distance stays away from 1 from below, then the oracle selector will prefer Q_n^*, and so will the cross-validation selector. Similarly, if is stays away from 1 from above, then the oracle selector will prefer Q_n and so will the cross-validation selector.

For concreteness and sake of illustration, assume $Q_n^* = Q_n(\epsilon_n)$ with $\epsilon_n = \arg\max_\epsilon P_n \log Q_n(\epsilon)$ and $Q_n(\epsilon)$ a fluctuation involving a known or fast converging g_n. In addition, assume that Q_n actually learns Q_0: i.e., for n converging to infinity, Q_n will converge to Q_0. We now make various

observations. Firstly, since ϵ_n is a maximum likelihood estimator according to a known parametric fluctuation, it will converge at a parametric rate to $\epsilon_{0n} = \arg\max_\epsilon P_0 \log Q_n(\epsilon)$, *if* the flucuation function is indeed known. If the parametric fluctuation function is identified by an estimator g_n, then the rate of convergence of $\epsilon_n - \epsilon_{0n}$ will be the worst of the parametric rate $1/\sqrt{n}$ and the rate at which g_n converges to its limit g. So typically (i.e., g_n will not converge faster than a parametric rate) this means that ϵ_n converges to ϵ_{0n} at the rate at which g_n converges to its limit.

In addition, a standard M-estimator analysis (i.e., Taylor expansion), based on the fact that ϵ_{0n} and $\epsilon_0 = 0 = \arg\max_\epsilon P_0 \log Q_0(\epsilon)$ solve their score equations, proves that ϵ_{0n} will converge to $\epsilon_0 = 0$ as a term involving a difference between Q_n and Q_0. Thus, we can then conclude that $\epsilon_n - \epsilon_0 = \epsilon_n - \epsilon_{0n} + (\epsilon_{0n} - \epsilon_0)$ $(\epsilon_0 = 0)$ will converge to zero at a rate that is the worst among the parametric rate, the rate of g_n and the rate of Q_n to Q_0.

We now note that the Kullback-Leibler dissimilarity, $P_0 \log Q_n(\epsilon_n)/Q_n$, behaves as ϵ_n^2 and thereby that $P_0 \log Q_n^*/Q_n$ will behave as ϵ_n^2. Therefore, it can be fully expected that the ratio $P_0 \log Q_n^*/Q_n / P_0 \log Q_n/Q_0$ will behave as the square of the worst of the rate at which g_n and Q_n converge, divided by the rate at which Q_n converges. Specifically, if Q_n converges slower than g_n, then this ratio will bounded away from zero and infinity, so that we can conclude that $d(Q_n^*, Q_0)/d(Q_n, Q_0)$ will behave as a random number between 0 and 1. Thus, in this case, the oracle selector will prefer Q_n^* above Q_n. On the other hand, if g_n converges slower than Q_n, then the ratio will converge to infinity so that $d(Q_n^*, Q_0)/d(Q_n, Q_0)$ will converge to infinity. Thus, in this case the oracle selector will prefer the initial estimator Q_n above Q_n^*.

The above reasoning can be formalized with regularity conditions. The clear qualitative conclusion is that the cross-validation selector will select for the targeting steps in which the directions are estimated at a rate good enough to not slow down the rate of convergence already achieved by the initial estimator.

4 Collaborative double robustness of estimating functions in CAR censored data models.

In this section we establish a new kind of collaborative robustness of the class of estimating functions in CAR-censored data models. The new result teaches us that the censoring mechanism required to obtain an unbiased estimating function at a mis-specified Q for the parameter of interest need not always condition on the whole full data structure. In fact, it teaches us that the better Q approximates Q_0 the less of an adjustment by full data random variables is necessary for the censoring mechanism to still obtain an unbiased estimating function for the parameter of interest. The precise collaborative property of $(Q, g_0(Q))$ required will be explicitly specified.

4.1 The formal collaborative robustness result.

The new form of double robustness we wish to establish is understood as follows. Consider an estimating function $D(\Psi(Q), G, Q)$ for the parameter of interest ψ_0 that is indexed by nuisance parameters (G_0, Q_0), and which is already known to establish the classical double robustness property: for any G under which ψ_0 is identifiable from $P_{Q_0,G}$, we have $E_0 D(\psi_0, G, Q) = 0$ if either $Q = Q_0$ or $G = G_0$ (van der Laan and Robins (2003)). Given a Q we are interested in the question under what conditional distribution $G_{0\delta}$ of censoring variable C, given a reduction $X(\delta)$ of X, will we still have $P_0 D(\psi_0, G_{0\delta}, Q) = 0$ and thereby that D is an unbiased estimating function for ψ at this mis-specified Q.

Firstly, we note that $P_0 D(\psi_0, G, Q) = P_0 D(\psi_0, G, Q) - D(\psi_0, G, Q_0) + P_0 D(\psi_0, G, Q_0)$, and the latter term is zero under any G that allows identifiability of ψ_0. Thus, it remains to determine for what $G_{0\delta}$ we will have $P_0 D(\psi_0, G_{0\delta}, Q) - D(\psi_0, G_{0\delta}, Q_0) = 0$.

By the general representation theorem for estimating functions that are orthogonal to nuisance scores of the full data model (Theorem 1.6, van der Laan and Robins (2003)), one can represent an estimating function $D(\psi_0, G, Q)$ as an Inverse Probability of Censoring Weighted Estimating function $D_{IPCW}(G, \psi_0)$ plus a function $D_{CAR}(Q, G)$ in the tangent space $T_{CAR}(G)$ of the censoring mechanism at G. The function $D_{CAR}(Q, G)$ is defined as the projection of $D_{IPCW}(G, \psi_0)$ on the tangent space $T_{CAR}(G) = \{h(O) :$

$E_G(h(O) \mid X) = 0\}$ of the censoring mechanism when only assuming coarsening at random, where this projection is carried out in the Hilbert space of all functions of O with mean zero and finite variance endowed with inner product the covariance operator $\langle f, g \rangle = Ef(O)g(O)$.

This teaches us that $P_0 D(\psi_0, G, Q) - D(\psi_0, G, Q_0) = P_0 D_{CAR}(Q, G) - D_{CAR}(Q_0, G)$, since the IPCW-difference equals zero. It also teaches us that for all Q we have that $D_{CAR}(Q, G)$ has conditional mean zero under G, given X. In addition, this same theorem also shows that $Q \to D_{CAR}(Q, G)$ is linear in Q. Therefore, it remains to show that $P_0 D_{CAR}(Q - Q_0, G) = 0$. Now, inspection of the proof that the conditional mean of $D_{CAR}(Q', G)$ under G equals zero for a Q' involves typically conditioning on a rich enough reduction of X so that a particular function indexed by Q' is fixed under the conditioning.

For example, for right censored data structures $O = (C, \bar{X}(C))$, $X(t)$ time dependent process, $\bar{X}(t) = \{X(s) : s \leq t\}$ representing the sample path up till time t, one can represent the projection of D_{IPCW} onto T_{CAR} as $D_{CAR}(Q, G) = \int H_{Q,G}(u, \bar{X}(u-)) dM_G(u)$, where

$$\begin{aligned} H_{Q,G}(u, \bar{X}(u-)) &= E_{Q,G}(D_{IPCW} \mid C = u, \bar{X}(u)) - E(D_{IPCW} \mid C \geq u, \bar{X}(u)) \\ dM_G(u) &= I(C = u) - I(C \geq u) d\Lambda_{C|X}(u \mid X), \end{aligned}$$

and $\Lambda_{C|X}$ is the cumulative hazard of C, given X. For details, we refer to van der Laan and Robins (2003), Chapter 3. Here $dM_G(u)$ is a Martingale satisfying $E(dM_G(u) \mid \bar{X}(u), C \geq u) = 0$. Due to the linearity of the conditional expectation operator, we have $D_{CAR}(Q - Q_0, G) = \int H_{Q-Q_0,G}(u, \bar{X}(u)) dM_G(u)$. By conditioning on $H_{Q-Q_0,G}(u, \bar{X}(u))$ within the integral, and using $E(dM_G(u) \mid \bar{X}(u), C \geq u) = 0$, it follows that $D_{CAR}(Q - Q_0, G)$ also has mean zero under a censoring mechanism s.t. $g_0(u \mid X)$ only depends on X through $H_{Q-Q_0,G}(u, \bar{X}(u))$. If Q approximates Q_0, this function $H_{Q-Q_0,G}$ will be shrunk to zero, so that less conditioning becomes necessary.

The following much simpler (but in essence making the same point) example helps to illustrate the general collaborative double robustness property. Suppose the observed censored data structure is $O = (W, \Delta, \Delta Y)$ and $X = (W, Y)$ is the full data random variable, where Δ is a censoring variable. Suppose one wishes to estimate $\psi_0 = E_0 Y$. The efficient influence curve is given by

$$D(\psi_0, \Pi_0, Q_0) = D_{IPCW}(\psi_0, \Pi_0) - D_{CAR}(Q_0, \Pi_0),$$

where

$$\begin{aligned} D_{IPCW}(\psi_0, \Pi_0) &= Y\frac{\Delta}{\Pi_0(W)} - \psi_0 \\ D_{CAR}(Q_0, \Pi_0) &= E(Y \mid \Delta = 1, W)\left(\frac{\Delta}{\Pi_0(W)} - 1\right), \end{aligned}$$

$\Pi_0(W) = P_0(\Delta = 1 \mid W)$ and $Q_0(W) = E_0(Y \mid W, \Delta = 1)$. Consider a Q. We are interested in the question under what conditional distribution $\Pi_{0\delta}$ of Δ, given a reduction $W(\delta)$ of W, will we still have $P_0 D(\psi_0, \Pi_{0\delta}, Q) = 0$ and thereby that D is an unbiased estimating function for ψ at this mis-specified Q. Firstly, we note that $P_0 D(\psi_0, \Pi, Q) = P_0 D(\psi_0, \Pi, Q) - D(\psi_0, \Pi, Q_0) + P_0 D(\psi_0, \Pi, Q_0)$, and the latter term is zero under any Π for which $P_0(\Pi(W) > 0) = 1$. Thus, it remains to determine for what $\Pi_{0\delta}$ $P_0 D(\psi_0, \Pi_{0\delta}, Q) - D(\psi_0, \Pi_{0\delta}, Q_0) = 0$.

This teaches us that $P_0 D(\psi_0, \Pi, Q) - D(\psi_0, \Pi, Q_0) = P_0 D_{CAR}(Q, \Pi) - D_{CAR}(Q_0, \Pi)$, since the IPCW-difference equals zero:

$$P_0 D(\psi_0, \Pi, Q) - D(\psi_0, \Pi, Q_0) = (Q - Q_0)(W)\left(\frac{\Delta}{\Pi_0(W)} - 1\right).$$

Note that we used here that $Q \to D_{CAR}(Q, \Pi)$ is linear in Q. Therefore, it remains to show that $P_0 D_{CAR}(Q - Q_0, \Pi) = 0$.

The proof that the conditional mean of $D_{CAR}(Q', \Pi)$ under Π equals zero for a Q' involves conditioning on a rich enough reduction of W so that $Q'(W)$ is captured by the conditioning: if $Q'(W)$ only depends on W through $W(\delta)$, then

$$EQ'(W)\left(\frac{\Delta}{\Pi_0(W(\delta))} - 1\right) = EQ'(W)\left(\frac{P_0(\Delta = 1 \mid W(\delta))}{\Pi_0(W(\delta))} - 1\right) = 0.$$

In particular, we have that the conditional mean of $D_{CAR}(Q - Q_0, \Pi_0)$, given $(Q - Q_0)(W)$, equals zero if $\Pi_0(W) = P(\Delta = 1 \mid Q - Q_0(W))$. This shows that if, for example, $(Q - Q_0)(W)$ only depends on one component W_1, then $P_0 D(\psi_0, \Pi_0, Q) = 0$ for $\Pi_0(W_1) = P_0(\Delta = 1 \mid W_1)$. That is, the better job Q does in approximating Q_0 the less inverse probability of missingness weighting is required to still obtain an unbiased estimating function for ψ_0.

We will now present the general result which can be applied to any CAR-censored data model as defined and studied in van der Laan and Robins (2003).

Theorem 1 *(***Collaborative Double Robustness of Efficient Influence Curve/Estimating Functions***)*

CAR-censored data model: *Let $O = \Phi(C, X) \sim P_0$ be a censored data structure with full data random variable $X \sim P_{X0}$, and censoring variable C with conditional probability distribution G_0 of C, given X. Assume G_0 satisfies the coarsening at random assumption. Let $g_0(C \mid X) = dG_0(C \mid X)$ a probability density of G_0 w.r.t. an appropriate dominating measure that satisfies coarsening at random itself. Let $\mathcal{M}$ denote the observed data model for P_0. Due to CAR, we have w.r.t. an appropriate dominating measure $dP_0(O) = Q_0(O)g_0(O \mid X)$, where $g_0(O \mid X)$ is only a function of O (by CAR), and Q_0 denotes the identifiable part of the full data distribution P_{X0}. (Here we abused notation to indicate that the conditional density of O, given X, is a deterministic function of the conditional density of C, given X, and, in fact, represents the identifiable part of the censoring mechanism G_0.) Let $\mathcal{Q}$ and $\mathcal{G}$ be models for Q_0 and G_0 which imply a model $\mathcal{M} = \{dP = Qg : Q \in \mathcal{Q}, G \in \mathcal{G}\}$ for P_0.*

Parameter of interest: *Let $\Psi : \mathcal{M} \to \mathbb{R}^d$ be pathwise differentiable parameter of interest and it is assumed that $\Psi(P_0) = \Psi^F(Q_0)$ is only a function of Q_0. Let $D^*(Q, G)$ be the efficient influence curve/canonical gradient of Ψ at $dP = Qg$.*

We make the following assumptions:

Augmented "PCW"-representation of efficient influence curve: *(PCW stands for Probability of Censoring Weighted) For each $Q \in \mathcal{Q}$, $G \in \mathcal{G}$,*

$$D^*(G, Q) = D_{h(G,Q)}(G, Q) = D_{h(G,Q),PCW}(G, \Gamma(Q)) + D_{h(G,Q),CAR}(G, Q'),$$

for mappings $(G, Q) \to h(G, Q)$, $(h, G, Q) \to D_{h,PCW}(G, \Gamma(Q))$, $(h, G, Q) \to D_{h,CAR}(G, Q)$, both defined on $\mathcal{H} \times \mathcal{G} \times \mathcal{Q}$, a parameter mapping Γ on $\mathcal{Q}$, and $(G, Q) \to Q'(G, Q)$.

(We refer to Theorem 1.6 in van der Laan and Robins (2003) for such a general representation of the efficient influence curve and, more generally, the orthogonal complement of the nuisance tangent space, where the CAR-components are elements of the tangent space T_{CAR} of G consisting of all functions of O with conditional mean zero, given X, under G. Under that representation, we have that $E_0 D_{h,PCW}(G_0, \gamma_0) = 0$ and $D_{h,CAR}(G_0, Q')$ has conditional mean zero, given X, for all Q'.)

Linearity of CAR-component: *$Q' \to D_{h,CAR}(G,Q')$ is linear on a set $\overline{\mathcal{Q}}'$ containing $\{Q'(G,Q) : G,Q\}$ in the sense that for all $h \in \mathcal{H}$, and all $Q_1, Q_2 \in \overline{\mathcal{Q}}'$*

$$D_{h,CAR}(G,Q_1') - D_{h,CAR}(G,Q_2') = D_{h,CAR}(G,Q_1' - Q_2').$$

Robustness for mis-specified censoring mechanism: *For all $Q_0 \in \mathcal{Q}_0$ and $G \in \mathcal{G}(Q_0) \subset \mathcal{G}$, where (e.g.,) $\mathcal{G}(Q_0)$ is defined as all censoring mechanisms G for which ψ_0 can be identified from $dP = dQ_0 g$, we have*

$$E_0 D_h(G,Q_0) = 0 \text{ for all } h \in \mathcal{H}.$$

Robustness of CAR-component: *For a reduction $X(\delta)$ of X (i.e., $X(\delta) = f(X,\delta)$ for some function f), let $G_{0\delta}$ be the conditional distribution of C, given $X(\delta)$.*

Let $\overline{\mathcal{Q}_\delta}'$ be a set within $\overline{\mathcal{Q}}'$ for which for each $\bar{Q}' \in \overline{\mathcal{Q}_\delta}'$

$$E_0 D_{h,CAR}(G_{0\delta}, \bar{Q}') = 0.$$

(Typically, one can select $\overline{\mathcal{Q}_\delta}'$ as all functions in $\overline{\mathcal{Q}}'$ that are only functions of X through $X(\delta)$.)

Let $\Gamma(Q) = \Gamma(Q_0)$ (typically implying $\Psi(Q) = \psi_0$), $G_{0\delta} \in \mathcal{G}(Q_0)$, and assume $Q' - Q_0' \in \overline{\mathcal{Q}_\delta}'$, where $Q' = Q'(G_{0\delta}, Q)$ and $Q_0' = Q'(G_{0\delta}, Q_0)$. Then

$$E_0 D^*(G_{0\delta}, Q) = 0.$$

We also have for all $G \in \mathcal{G}(Q_0)$

$$E_0 D^*(G, Q_0) = 0.$$

Proof. Suppose $\Gamma(Q) = \Gamma(Q_0)$ and $Q' - Q_0' \in \overline{\mathcal{Q}_\delta}'$. Let $G_0^* = G_{0\delta}$ be the conditional distribution of C, given $X(\delta)$, and assume it is an element of $\mathcal{G}(Q_0)$.

By the "Augmented 'PCW'-representation of efficient influence curve" assumption, we have

$$E_0 D^*(G_0^*, Q) = E_0 D_h(G_0^*, Q)$$

for some $h \in \mathcal{H}$. Thus,

$$\begin{aligned} E_0 D^*(G_0^*, Q) &= E_0 D_h(G_0^*, Q) \\ &= E_0\{D_h(G_0^*, Q) - D_h(G_0^*, Q_0)\} + E_0 D_h(G_0^*, Q_0). \end{aligned}$$

By the assumption that $G_0^* \in \mathcal{G}(Q_0)$, it follows that the last term $E_0 D_h(G_0^*, Q_0) = 0$.

By the "PCW-representation" assumption we have

$$\begin{aligned} E_0\{D_h(G_0^*, Q) - D_h(G_0^*, Q_0)\} &= E_0\{D_{h,PCW}(G_0^*, \Gamma(Q)) - D_{h,PCW}(G_0^*, \Gamma(Q_0))\} \\ &\quad + E_0\{D_{h,CAR}(G_0^*, Q'(Q, G_0^*)) - D_{h,CAR}(G_0^*, Q'(Q_0, G_0^*))\}. \end{aligned}$$

By the assumption that $\Gamma(Q) = \Gamma(Q_0)$, the first term equals zero. By the "linearity of CAR-component"-assumption we have that the last term equals:

$$E_0\{D_{h,CAR}(G_0^*, Q') - D_{h,CAR}(G_0^*, Q_0')\} = E_0 D_{h,CAR}(G_0^*, Q' - Q_0'),$$

where $Q' = Q'(G_0^*, Q)$ and $Q_0' = Q'(G_0^*, Q_0)$.

We assumed that $Q' - Q_0' \in \overline{\mathcal{Q}_\delta}'$. Thus, by the "Robustness of CAR-component"-assumption we have that

$$E_0 D_{h,CAR}(G_0^*, Q' - Q_0') = 0.$$

This proves $E_0 D^*(G_0^*, Q) = 0$. □

For the sake of explicit illustration, we will now explicitly establish the collaborative double robustness of the efficient influence curve estimating function in two additional examples. These results are also corollaries of the above general Theorem 1.

4.2 Example I: Marginal causal effect in nonparametric model.

Let $O = (W, A, Y) \sim P_0$ and let the model $\mathcal{M}$ be nonparametric. Let $\Psi(P_0) = E_0(E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W))$ be the parameter of interest. This parameter can be referred to as a variable importance of variable A or, under additional causal inference assumptions, it can be interpreted as a causal effect.

The probability distribution of O can be factored as

$$dP_0(W, A, Y) = Q_{01}(W) Q_{02}(Y \mid A, W) dG_0(A \mid W),$$

where each factor represents a density w.r.t. an appropriate dominating measure of the conditional distribution. Let $W(\delta) \subset W$ be a subset of W indexed by an index δ.

The efficient influence curve of Ψ at $P = (Q, G)$ can be represented as

$$D^*(Q,G)(O) = h(G)(A,W)(Y - Q_2(A,W)) + Q_2(1,W) - Q_2(0,W) - \Psi(Q),$$

where $Q_2(A,W) = E_{Q_2}(Y \mid A, W)$ denotes the conditional mean of Y, given A, W, under $Q = (Q_1, Q_2)$, and $h(G)(A,W) = A/g(1 \mid W) - (1-A)/g(0 \mid W)$, and $g(1 \mid W) = P_G(A = 1 \mid W)$.

We have the following double robustness result.

Theorem 2 *Let $dP_0 = Q_0 dG_0$ be the distribution of $O = (W, A, Y)$ and let the model for P_0 be nonparametric.*

Let $\Psi(Q_0) = E_{Q_{01}}\{E_{Q_{02}}(Y \mid A = 1, W) - E_{Q_{02}}(Y \mid A = 0, W)\}$ be the parameter on this model, where it is assumed that it is identifiable from P_0. The efficient influence curve of Ψ at $P = (Q, G)$ is given by

$$D^*(Q,G)(O) = h(G)(A,W)(Y - Q_2(A,W)) + Q_2(1,W) - Q_2(0,W) - \Psi(Q),$$

where $Q_2(A,W) = E_Q(Y \mid A, W)$ denotes the conditional mean of Y, given A, W, under $Q = (Q_1, Q_2)$.

Assume

$$(Q_{02} - Q_2)(A,W) = E_{Q_0}(Y - Q_2(A,W) \mid A, W) = f_0(A, W(Q))$$

is only a function of $A, W(Q)$ for a $W(Q) = \Phi(Q_2, W)$ for some mapping Φ: i.e., $W(Q)$ denotes a reduction or subset of the full vector random variable W indexed by Q.

Let $dG_0(Q)$ be the conditional distribution of A, given $W(Q)$. If $\Psi(Q) = \Psi(Q_0)$, then

$$E_{P_0} D^*(Q, G_0(Q)) = 0.$$

Or, equivalently, if we represent $D^*(Q,G)$ as $D^*(\Psi(Q), Q, G)$, then

$$E_{P_0} D^*(\psi_0, Q, G_0(Q)) = 0.$$

We also have: If $Pr(P_G(A = 0 \mid W) * P_G(A = 1 \mid W) > 0) = 1$, then

$$E_{P_0} D^*(Q_0, G) = 0,$$

or equivalently,

$$E_{P_0} D^*(\psi_0, Q_0, G) = 0.$$

Proof. The last statement is easy and well known (e.g., van der Laan and Robins (2003)). The first statement needs to be proved, or can be derived as a corollary of Theorem 1. Note, if $\Psi(Q) = \psi_0$, then

$$E_0 D^*(Q, G_0(Q)) = E_0 h(G_0)(A, W(Q))(Y - Q(A, W)) + Q(1, W) - Q(0, W) - \psi_0.$$

If $E_0(Y - Q(A, W) \mid A, W) = f_0(A, W(Q))$ is only a function of $A, W(Q)$, then it follows by first taking the conditional mean, given A, W, and then taking the mean of A, given $W(Q)$,

$$\begin{aligned} E_0 D^*(Q, G_0(Q)) &= E_0 h(G_0)(A, W(Q)) f_0(A, W(Q)) \\ &\quad + Q(1, W) - Q(0, W) - \psi_0 \\ &= E_0 f_0(1, W(Q)) - f_0(0, W(Q)) + Q(1, W) - Q(0, W) \\ &\quad - \psi_0. \end{aligned}$$

Now, note that $f_0(A, W(Q)) = Q_0(A, W) - Q(A, W)$, which proves that the latter quantity equals zero.

□

The implication of this result is that, given an estimate Q of Q_0, we only need to estimate $G_0(Q)$ as a parameter of the complete conditional distribution G_0. Thus, if Q already succeeds in explaining most of the true regression $E_0(Y \mid A, W)$, then only little inverse weighting with $G_0(Q) = P(A = \cdot \mid W(Q))$ remains to be done. That is, the amount of inverse weighting required to obtain a consistent estimator of the causal effect ψ_0 can be adapted to the degree to which the true regression is fit by Q.

4.3 Example II: Semiparametric regression.

Let $O = (W, A, Y) \sim P_0$. Assume the model $E_0(Y \mid A, W) - E_0(Y \mid A = 0, W) = A\beta_0 V$ for some $V \subset W$. If the variance of Y, given A, W, only depends on W, then the efficient score of β_0 at P_0 can be represented as

$$D^*(\Pi_0, \theta_0, \beta_0)(O) = (A - \Pi_0(W))(Y - A\beta_0 V - \theta_0(W)),$$

where $\Pi_0(W) = E_0(A \mid W)$, and $\theta_0(W) = E_0(Y \mid A = 0, W)$. For the sake of illustration we will use this simpler representation, but the same double robustness applies to the general efficient influence curve representation as (e.g.) presented in van der Laan and Robins (2003).

Theorem 3 *Suppose* $E_0(Y - A\beta_0 V - \theta(W) \mid A, W) = f_0(W(\theta))$ *for some function* f_0 *of* $W(\theta)$ *where* $W(\theta) = \Phi(W, \theta)$ *is function of* W *and* θ*. Note that this states that* $\theta_0(W) - \theta(W) = f_0(W(\theta))$ *is only a function of a reduction* $W(\theta)$ *of* W*. Let* $\Pi_0(\theta)(W) = E_0(A \mid W(\theta))$*. Then*

$$E_0 D^*(\Pi_0(\theta), \theta, \beta_0) = 0$$

We also have

$$E_0 D^*(\Pi, \theta_0, \beta_0) = 0$$

Proof. Only the first robustness result needs to be proved. First take the conditional mean, given A, W, which results in the term $E_0(A - \Pi_0(\theta)(W(\theta)))f_0(W(\theta))$. Subsequently, we take the conditional mean, given $W(\theta)$, which proves it equals zero. □

5 Asymptotics of collaborative TMLE.

The collaborative targeted maximum likelihood estimator Q_n^* equals a k_n-th collaborative targeted maximum likelihood estimator, and thereby equals a targeted maximum likelihood estimator with a starting estimator Q_n (the $k_n - 1$-th collaborative targeted maximum likelihood estimator), and the censoring mechanism estimator $g_n = g_{n\delta_n}$ as selected in the k_n-step, given a collection of candidate estimators $g_{n\delta}$. Thus, just like the targeted maximum likelihood estimator, the collaborative targeted maximum likelihood estimator $\psi_n = \Psi(Q_n^*)$ of ψ_0 solves the efficient influence curve estimating equation

$$0 = P_n D^*(Q_n^*, g_n, \psi_n).$$

For simplicity, we will make the assumption that the efficient influence curve at a $P_{Q,g}$ can be represented as an estimating function in ψ. However, the theorem in this section can be generalized to any efficient influence curve $D^*(Q, g)$ identified by a data generating distribution $P_{Q,g}$.

Since Q_n^* is just a maximum likelihood based estimator within a model, it is a reasonable assumption that Q_n^* converges to some element Q in the model for Q_0. In addition, $g_{n\delta}$ converges to some $g_{0\delta}$ for each δ ranging over a finite set. If each extra iteration corresponds with a unique targeting step, and we use the cross-validated log-likelihood to select the number of iterations k_n in the C-TMLE, then this selector will in most cases result in a

data adaptive choice δ_n so that the corresponding complexity-measure $d(\delta_n)$ converges either to infinity (i.e., resulting in most nonparametric estimator consistent for $G_0(\cdot \mid X)$) or some upper limit (i.e., it stops short of selecting the most nonparametric estimator). In these cases one expects that, as n converges to infinity, then g_n converges to a fixed $g_{0\delta_0}$ representing the limit of a $g_{n\delta_0}$, not necessarily equal to the conditional distribution given the full X. For notational convenience, we will denote this limit with g_0.

In such a case one will have that

$$P_0 D^*(Q, g_{0\delta}, \psi_0) = 0 \text{ for } d(\delta) \geq d(\delta_0),$$

and thereby that, in particular,

$$0 = P_0 D^*(Q, g_0, \psi_0),$$

which will be the fundamental assumptions for asymptotic normality of the C-TMLE.

Theorem 4 *Let $(Q, g, \psi) \to D^*(Q, g, \psi)$ be a well defined function that maps any possible $(Q, g, \Psi(Q))$ into a function of O. Let $O_1, \ldots, O_n \sim P_0$ be i.i.d, and let P_n be the empirical probability distribution. Let $Q \to \Psi(Q)$ be a d-dimensional parameter, where $\psi_0 = \Psi(Q_0)$ is the parameter value of interest. In the following template Q_n^* represents the collaborative targeted maximum likelihood estimator, but it can be any estimator.*

Assume

- $0 = P_0 D^*(Q, g_0, \psi_0) = P_n D^*(Q_n^*, g_n, \psi_n)$, *where* $\psi_n = \Psi(Q_n^*)$.
- $$P_0(D^*(Q_n^*, g_n, \psi_n) - D^*(Q, g_0, \psi_0))^2 \to 0 \text{ in probability,}$$
  as $n \to \infty$. And the same applies if one or two of the triplets (Q_n^*, g_n, ψ_n) is replaced by its limit (Q, g_0, ψ_0).
- $c_0 = -d/d\psi_0 P_0 D^*(Q, g_0, \psi_0)$ *exists and is invertible.*
- $\{D^*(Q, g, \Psi(Q)) : Q, g\}$ *is P_0-Donsker, where (Q, g) vary over sets that contain (Q_n^*, g_n), (Q^*, g_n), (Q_n^*, g) with probability tending to 1.*
- *Define the mapping $Q \to \Phi_1(Q) \equiv P_0 D^*(Q, g_0, \psi_0)$. Assume $\Phi_1(Q_n^*) - \Phi_1(Q) = (P_n - P_0) IC_Q + o_P(1/\sqrt{n})$ for some mean zero function $IC_Q \in L_0^2(P_0)$.*

- *Define the mapping* $g \rightarrow \Phi_2(g) \equiv P_0 D^*(Q, g, \psi_0)$. *Assume* $\Phi_2(g_n) - \Phi_2(g_0) = (P_n - P_0)IC_{g_0} + o_P(1/\sqrt{n})$ *for some mean zero function* $IC_{g_0} \in L_0^2(P_0)$.

- *Define second order term*

  $$R_{n1} = P_0\{D^*(Q_n^*, g_n, \psi_n) - D^*(Q, g_n, \psi_n)\} - \{D^*(Q_n^*, g_0, \psi_0) - D^*(Q, g_0, \psi_0)\},$$

  and assume $R_{n1} = o_P(1/\sqrt{n})$. *Note* R_{n1} *is a second order term involving difference between* $Q_n^* - Q$ *and* $g_n - g_0$.

- *Define second order term*

  $$R_{n2} = P_0\{D^*(Q, g_n, \psi_n) - D^*(Q, g_0, \psi_n)\} - \{D^*(Q, g_n, \psi_0) - D^*(Q, g_0, \psi_0)\},$$

  and assume $R_{n2} = o_P(1/\sqrt{n})$. *Note* R_{n2} *is a second order term involving difference between* $g_n - g_0$ *and* $\psi_n - \psi_0$.

Then, ψ_n *is asymptotically linear estimator of* ψ_0 *with influence curve*

$$IC(P_0) = c_0^{-1} \left\{ D^*(Q, g_0, \psi_0) + IC_Q + IC_{g_0} \right\}.$$

That is,

$$\psi_n - \psi_0 = (P_n - P_0)IC(P_0) + o_P(1/\sqrt{n}).$$

In particular, $\sqrt{n}(\psi_n - \psi_0)$ *converges in distribution to a multivariate normal distribution with mean zero and covariance matrix* $\Sigma_0 = E_0 IC(P_0) IC(P_0)^\top$.

Proof: The principle equations are $0 = P_n D^*(Q_n, g_n, \psi_n)$ and $P_0 D^*(Q, g_0, \psi_0) = 0$. So, we have

$$P_0 D^*(Q, g_0, \psi_n) - D^*(Q, g_0, \psi_0) = -\left\{ P_n D^*(Q_n, g_n, \psi_n) - P_0 D^*(Q, g_0, \psi_n) \right\}.$$

Let $c_0 = -\frac{d}{d\psi_0} P_0 D^*(Q, g_0, \psi_0)$. Then,

$$\begin{aligned} c_0(\psi_n - \psi_0) + o(\mid \psi_n - \psi_0 \mid) \;\; &= \;\; (P_n - P_0)D^*(Q, g_0, \psi_n) \\ &+ P_n\{D^*(Q_n, g_n, \psi_n) - D^*(Q, g_n, \psi_n)\} + P_n\{D^*(Q, g_n, \psi_n) - D^*(Q, g_0, \psi_n)\}. \end{aligned}$$

We denote the three terms on the right with I,II and III, and deal with them separately below. In this proof, we will refer to the second assumption as consistency condition (on the estimated influence curve).

I: By the Donsker condition, and consistency condition, we have

$$(P_n - P_0)\{D^*(Q, g_0, \psi_n) - D^*(Q, g_0, \psi_0)\} = o_P(1/\sqrt{n}).$$

Thus, we obtain $(P_n - P_0)D^*(Q, g_0, \psi_0) + o_P(1/\sqrt{n})$ as first term approximation.
II: We have

$$\begin{aligned} P_n\{D^*(Q_n, g_n, \psi_n) - D^*(Q, g_n, \psi_n)\} &= (P_n - P_0)\{D^*(Q_n, g_n, \psi_n) - D^*(Q, g_n, \psi_n)\} \\ &\quad + P_0\{D^*(Q_n, g_n, \psi_n) - D^*(Q, g_n, \psi_n)\}. \end{aligned}$$

The first term is $o_P(1/\sqrt{n})$ by Donsker class condition, and consistency condition at Q_n, g_n, ψ_n. We also have

$$P_0\{D^*(Q_n, g_n, \psi_n) - D^*(Q, g_n, \psi_n)\} = P_0\{D^*(Q_n, g_0, \psi_0) - D^*(Q, g_0, \psi_0)\} + R_{n1},$$

where

$$\begin{aligned} R_{n1} &= P_0\{D^*(Q_n, g_n, \psi_n) - D^*(Q, g_n, \psi_n) - D^*(Q_n, g_0, \psi_0) - D^*(Q, g_0, \psi_0)\} \\ &= o_P(1/\sqrt{n}), \end{aligned}$$

by assumption. R_{n1} is a second order term involving $Q_n - Q$ and $(g_n, \psi_n) - (g_0, \psi_0)$. Thus the second term equals $P_0\{D^*(Q_n, g_0, \psi_0) - D^*(Q, g_0, \psi_0)\}$. This equals $\Phi_1(Q_n) - \Phi_1(Q)$. We assumed that $\Phi_1(Q_n) - \Phi_1(Q) = (P_n - P_0)IC_Q + o_P(1/\sqrt{n})$. Thus, the second term equals $(P_n - P_0)IC_Q + o_P(1/\sqrt{n})$.

III:

We have

$$\begin{aligned} P_n D^*(Q, g_n, \psi_n) - D^*(Q, g_0, \psi_n) &= (P_n - P_0)D^*(Q, g_n, \psi_n) - D^*(Q, g_0, \psi_n) \\ &\quad + P_0 D^*(Q, g_n, \psi_n) - D^*(Q, g_0, \psi_n). \end{aligned}$$

The first term is $o_P(1/\sqrt{n})$ by Donsker class condition, and consistency condition at Q_n, g_n, ψ_n. We also have

$$P_0 D^*(Q, g_n, \psi_n) - D^*(Q, g_0, \psi_n) = P_0 D^*(Q, g_n, \psi_0) - D^*(Q, g_0, \psi_0) + R_{n2},$$

where

$$R_{n2} = P_0 D^*(Q, g_n, \psi_n) - D^*(Q, g_0, \psi_n) - D^*(Q, g_n, \psi_0) - D^*(Q, g_0, \psi_0) = o_P(1/\sqrt{n}),$$

by assumption. Thus the third term equals $P_0D^*(Q,g_n,\psi_0) - D^*(Q,g_0,\psi_0)$. This equals $\Phi_2(g_n) - \Phi_2(g_0)$. We assumed that $\Phi_2(g_n) - \Phi_1(g_0) = (P_n - P_0)IC_{g_0} + o_P(1/\sqrt{n})$. Thus, the third term equals $(P_n - P_0)IC_{g_0} + o_P(1/\sqrt{n})$.

We can thus conclude that

$$\psi_n - \psi_0 = (P_n - P_0)c_0^{-1}\{D^*(Q,g_0,\psi_0) + IC_Q + IC_{g_0}\} + o_P(\mid \psi_n - \psi_0 \mid) + o_P(1/\sqrt{n}).$$

This implies $\mid \psi_n - \psi_0 \mid = O_P(1/\sqrt{n})$, and thereby the stated asymptotic linearity. □

5.1 Irregular C-TMLE and super efficiency.

Due to the particular way the g_n is constructed in response to Q_n, it is easily argued that the collaborative targeted MLE can be an irregular estimator and can be super efficient by achieving an asymptotic variance that is smaller than the variance of the efficient influence curve. In particular, our previous arguments showed that if the initial estimator is a maximum likelihood estimator according to a correctly specified parametric model, then g_n will avoid nonparametric fits, thereby staying away from estimating the g_0 that would result in an efficient estimator in first order. In these cases we observed super efficiency in our simulations. Indeed, in this case the influence curve will now be of the form $D^*(Q,g_0,\psi_0) + IC_Q$, where IC_Q is an influence curve that is a product of a delta-method applied to Q_n as an estimator of Q_0.

6 Penalized targeted log-likelihood criterion.

Consider candidate (e.g., collaborative) targeted maximum likelihood estimators $P_n \rightarrow \hat{P}^*_\delta(P_n)$ of the true probability distribution of the data $P_0 \in \mathcal{M}$, targeting a parameter $\psi_0 = \Psi(P_0)$, indexed by δ. Our proposed criterion for selecting δ is

$$\delta_n = \underset{\delta}{\text{argmax}}\ L_{CV}(P_n)(\delta) + L^*(P_n)(\delta) - MSE(P_n)(\delta),$$

or

$$\delta_n = \underset{\delta}{\text{argmax}}\ L^*_{CV}(P_n)(\delta) - MSE(P_n)(\delta).$$

These terms will be specified below.

6.1 The cross-validated log-likelihood of (T)MLE.

Given a particular dominating measure, we define the cross-validated log-likelihood at the δ-specific targeted maximum likelihood estimator:

$$L_{CV}(P_n)(\delta) = E_{B_n} P^1_{n,B_n} \log \frac{d\hat{P}^*_\delta(P^0_{n,B_n})}{d\mu},$$

where $B_n \in \{0,1\}^n$ is a random binary vector of length n defining a split of the sample into a training sample $\{i : B_n(i) = 0\}$ and validation sample $\{i : B_n(i) = 1\}$, whose empirical distributions are denoted with P^0_{n,B_n} and P^1_{n,B_n}, respectively.

We typically restrict attention to V-fold cross-validation so that B_n has only V realizations, even though the choice of cross-validation scheme is user supplied. In V-fold cross-validation these V realizations are identified by a split of the original sample into V equal size sub-samples, by defining the validation sample as one of these V subgroups, and the training sample as the complement. Note that evaluation of this cross-validated log-likelihood requires computation of the targeted maximum likelihood estimator $P_n \rightarrow \hat{P}^*_\delta(P_n)$ for each of the V training samples.

One could select δ as the maximizer of the cross-validated log-likelihood at the TMLE.

As discussed previously, by finite sample and asymptotic oracle results established for this likelihood-based cross-validation selector (van der Laan et al. (2004)), this results in an excellent estimator of the actual distribution P_0 optimally trading off bias and variance w.r.t. the target P_0 w.r.t the Kullback-Leibler measure of dissimilarity between two densities/data generating distributions. Since the targeted maximum likelihood estimators are all targeted towards ψ_0, in many applications this cross-validated targeted log-likelihood criterion works well.

Therefore this log-likelihood term provides an excellent basis of our proposed targeted criterion that will drive the selector towards a fit of the true P_0 and thereby the target parameter ψ_0. Our sole motivation for the proposed additional penalty terms is to make the criterion more targeted towards ψ_0.

We also note that if the density of O factorizes, $dP_0 = dQ_0 dG_0$, in two factors dQ and dG, and the parameter of interest $\Psi(P)$ only depends on Q, then we replace this cross-validated log-likelihood by the relevant term of the

log-likelihood:

$$L_{CV}(P_n)(\delta) = E_{B_n} P^1_{n,B_n} \log \frac{d\hat{Q}^*_\delta(P^0_{n,B_n})}{d\mu}.$$

Here $P_n \to \hat{Q}^*_\delta(P_n)$ denotes the targeted maximum likelihood estimator of the relevant factor Q_0, where $d\hat{P}^*_\delta(P_n) = d\hat{Q}^*_\delta(P_n) d\hat{G}^*_\delta(P_n)$. This targeted maximum likelihood estimator of Q_0 will still depend on the targeted maximum likelihood estimator of G_0, since the updating step in the TMLE-algorithm will depend on this estimator.

The term we present in the next subsection will measure separately the gain in the log-likelihood due to the targeting step in the targeted maximum likelihood algorithm. If one uses this term in the criterion to select δ, then one can may decide to use the the cross-validated log-likelihood at the δ-specific initial estimator $\hat{P}_\delta$ (instead of at the corresponding C-TMLE):

$$L_{CV}(P_n)(\delta) = E_{B_n} P^1_{n,B_n} \log \frac{d\hat{P}_\delta(P^0_{n,B_n})}{d\mu}.$$

6.2 The empirical log-likelihood increase due to targeting.

The increase of the log-likelihood during the targeted maximum likelihood updating algorithm corresponds with fitting the parameter of interest. Therefore, this increase is all about bias reduction for the parameter of interest and a choice of δ that results in a large bias reduction should be rewarded.

Therefore, we may add the following term to the cross-validated log-likelihood of the TMLE:

$$L^*(P_n)(\delta) = P_n \log \frac{d\hat{P}^*_\delta(P_n)}{d\hat{P}_\delta(P_n)},$$

where $\hat{P}_\delta(P_n)$ is the initial δ-specific estimator of P_0 (i.e., the first stage estimator) and $\hat{P}^*_\delta(P_n)$ is the (iterative) δ-specific targeted maximum likelihood estimator of P_0 taking $\hat{P}_\delta(P_n)$ as initial estimator.

This term may be appropriate for first step collaborative targeted maximum likelihood estimators, but might be subject to over-fitting for selecting among k-th step C-TMLE in which k can be large. Therefore, in the latter

case, we propose to replace the k-th step C-TMLE in the above term by the TMLE using the k-th censoring/nuisance parameter estimator with initial estimator $\hat{P}_\delta(P_n)$ (thus dropping the previous fluctuations). In this way, the above term measures a bias gain due to the targeted maximum likelihood algorithm using the last selected censoring mechanism estimator.

In our nonparametric causal effect example one can show that this increase in log-likelihood due to the targeted maximum likelihood algorithm (now only requiring one step) equals the squared difference of the substitution estimator $\Psi(\hat{P}_\delta(P_n))$ and the targeted maximum likelihood substitution estimator $\Psi(\hat{P}^*_\delta(P_n))$, scaled appropriately.

Again, if $dP_0 = dQ_0 dG_0$ factorizes in two factors and the parameter of interest $\Psi(P_0)$ only depends on Q_0, then we replace this targeted log-likelihood increase by the relevant term:

$$L^*(P_n)(\delta) = P_n \log \frac{d\hat{Q}^*_\delta(P_n)}{d\hat{Q}_\delta(P_n)}.$$

This targeted log-likelihood increase term make our proposed selector clearly more targeted in censored data and causal inference semi-parametric models: For example, if there are no good choices of δ w.r.t. to $\hat{Q}_\delta(P_n)$ as an estimator of Q_0, but there is a choice of δ that results in great bias reduction for the target parameter due to the targeted maximum likelihood algorithm, then our selector will select the latter δ.

6.3 Combining the log-likelihood and targeted increase of log-likelihood in one criterion.

Here we propose a modification of the cross-validated log-likelihood at a TMLE, which also fully captures the targeted increase of the log-likelihood as measured by $L^*(P_n)$ as defined in previous subsection. This new term which we will denote with $L^*_{CV}(P_n)$ can replace $L_{CV}(P_n) + L^*(P_n)$ in our final criterion.

Define

$$L_{CV,\delta}(\epsilon) \equiv E_{B_n} P^1_{n,B_n} \log \frac{d\hat{P}_\delta(P^0_{n,B_n})(\epsilon)}{d\mu}.$$

Let

$$\epsilon^0_n = \underset{\epsilon}{\text{argmax}}\, L_{CV,\delta}(\epsilon).$$

This value can now be used to define a first step targeted MLE update $\hat{P}_\delta(P^0_{n,B_n})^1 = \hat{P}_\delta(P^0_{n,B_n})(\epsilon^0_n)$ on the training sample P^0_{n,B_n} and corresponding $L^0_{CV}(\delta) = L_{CV,\delta}(\epsilon^0_n)$. In the case that this process converges in one step, we would use $L^0_{CV}(\delta)$ in the criterion for selecting δ. Note that, it measures the increase of the likelihood of the validation observations due to fluctuating the training sample based initial estimator.

In general, we iterate by defining

$$L^1_{CV,\delta}(\epsilon) \equiv E_{B_n} P^1_{n,B_n} \log \frac{d\hat{P}_\delta(P^0_{n,B_n})^1(\epsilon)}{d\mu}.$$

We now define a process that maps an initial $L_{CV,\delta}(\epsilon)$ and corresponding maximum $L_{CV,\delta}(\epsilon^0_n)$ into an updated $L^1_{CV,\delta}(\epsilon)$ and corresponding $L^1_{CV,\delta}(\epsilon^1_n)$, where $\epsilon^1_n = \underset{\epsilon}{\text{argmax}}\ L^1_{CV,\delta}(\epsilon)$. This process can now be iterated till convergence and the final value of $L^k_{CV,\delta}(\epsilon^k_n)$ is used as a criterion for δ.

We now use as single cross-validated targeted maximum likelihood criterion:

$$L^*_{CV}(P_n)(\delta) = L^k_{CV,\delta}(\epsilon^k_n).$$

If the density of O factorizes, $dP_0 = dQ_0 dG_0$, then the log-likelihood terms are replaced by $\log(d\hat{Q}_\delta(P^0_{n,B_n})^k(\epsilon)/d\mu)$ for the appropriate dominating measure $d\mu$. In addition, if the epsilon-extension $Q(\epsilon)$ is indexed by G_0, then we suggest that it is appropriate to use the estimator $\hat{G}_\delta(P_n)$ on the whole sample, instead of applying this estimator to the training samples P^0_{n,B_n}. That is, the cross-validation does not need to be strictly applied to $\hat{G}_\delta$ since this estimator is based on an orthogonal factor in the likelihood.

6.4 Variance of targeted maximum likelihood estimator relative to its δ-limit.

If the target parameter cannot be reasonably identified from the data the log-likelihood terms above will not be sensitive enough to such a singularity: in fact, on many occasions this just means that the targeted maximum likelihood algorithm will be ineffective (i.e., the maximum likelihood fluctuations get too noisy) so that in essence the log-likelihood of the initial estimator drives the selection.

Therefore it is crucial that the log-likelihood terms are penalized by a term that blows up (in the negative direction) for δ-values for which the

variance (or bias, addressed in next subsection) of the targeted maximum likelihood estimator $\Psi(\hat{P}^*_\delta(P_n))$ relative to its limit $\psi_0(\delta) = \Psi(\hat{P}^*_\delta(P_0))$ gets large. Since we can derive the influence curve of the targeted maximum likelihood estimator $\Psi(\hat{P}^*_\delta(P_n))$ as an estimator of $\psi_0(\delta)$, this variance can be estimated with the variance of this influence curve at this targeted maximum likelihood estimator $\hat{P}^*_\delta(P_n)$. As follows from the study of TMLE in van der Laan and Rubin (2006) one can often use as influence curve the efficient influence curve $D^*(P)$, at $P = \hat{P}^*_\delta(P_n)$, of the parameter $\Psi : \mathcal{M} \to \mathbb{R}^d$.

We first define the cross-validated covariance matrix for the estimator $\hat{\Psi}(\hat{P}^*_\delta(P_n))$:

$$\frac{\Sigma(P_n)(\delta)}{n} = \frac{1}{n} E_{B_n} P^1_{n,B_n} \left\{ D^*(\hat{P}^*_\delta(P^0_{n,B_n})) D^*(\hat{P}^*_\delta(P^0_{n,B_n}))^\top \right\}.$$

For example, if the target parameter is 1-dimensional (i.e., $d = 1$), then we have

$$\frac{\sigma^2(P_n)(\delta)}{n} = \frac{1}{n} E_{B_n} P^1_{n,B_n} \left\{ D^*(\hat{P}^*_\delta(P^0_{n,B_n})) \right\}^2.$$

In a next subsection we discuss how this covariance matrix can be used to construct a MSE term for our penalized log-likelihood. For example, one can define the variance term of the MSE in our penalized log-likelihood criterion, as

$$\sigma^2(P_n)(\delta) = a \Sigma(P_n)(\delta) a^\top,$$

for a user supplied vector a, so that $\sigma^2(P_n)/n$ represents the variance estimate of the estimator of $a^\top \psi_0(\delta)$.

Our proposal will actually have the form

$$\sigma^2(P_n)(\delta) = \sum_{j=1}^{d} a_j^\top \Sigma(P_n)(\delta) a_j, \tag{2}$$

where a_j are the row vectors of the square root of a user supplied matrix such as the inverse of the the correlation matrix of $\Sigma(P_n)(\delta)$.

6.5 Bias of targeted maximum likelihood estimator relative to its δ-limit.

By the same argument, we wish to estimate the bias of the targeted maximum likelihood estimator $\Psi(\hat{P}^*_\delta(P_n))$ relative to its limit $\psi_0(\delta)$. For example, this

could be done with the bootstrap:

$$E_{P_n}\left\{\Psi(\hat{P}^*_\delta(P^\#_n)) - \Psi(\hat{P}^*_\delta(P_n))\right\},$$

where $P^\#_n$ represents the empirical distribution of a bootstrap sample $O^\#_1, \ldots, O^\#_n$ from the empirical distribution P_n. However, this would be much too computer intensive in many applications in which the targeted maximum likelihood estimator involves data adaptive model or algorithm selection. By noting that a bootstrap sample corresponds on average with 2/3 of the n observations, the following analogue bias estimate can be viewed as an approximation of this bootstrap bias that only requires 3 times applying the targeted maximum likelihood estimator to a sample of size $n * 2/3$:

$$B(P_n)(\delta) = E_{B_{n3}}\left\{\Psi(\hat{P}^*_\delta(P^0_{n,B_n})) - \Psi(\hat{P}^*_\delta(P_n)\right\},$$

where B_{n3} denotes the 3-fold cross-validation scheme.

If $d = 1$, then we will add to the variance term in the previous section the squared bias $B(P_n)^2$ to create a MSE-term. If $d > 1$, then in our proposal below we will construct an appropriate function of $B(P_n)$ representing the analogue of the variance term (2):

$$b(P_n)^2(\delta) \equiv \sum_j (a_j^\top B(P_n)(\delta))^2.$$

Additional rationale behind bias term: To provide further understanding of this kind of bias estimate $B(P_n)$, we note the following. Let $\hat{\Psi}(P_n)$ be an estimator of its target $\hat{\Psi}(P_0)$, where it plays the role of the δ-specific targeted maximum likelihood estimator $\Psi(\hat{P}^*_\delta(P_n))$. The fundamental assumption allowing statistical inference for $\hat{\Psi}(P_0)$ is the assumption of asymptotic linearity:

$$\hat{\Psi}(P_n) - \hat{\Psi}(P_0) = (P_n - P_0)D(P_0) + R(P_n), \tag{3}$$

where $D(P_0)$ is the influence curve of the estimator, and $R(P_n)$ is the remainder. The asymptotic linearity assumption now assumes that $R(P_n) = o_P(1/\sqrt{n})$.

The representation (3) of the mapping $P_n \to \hat{\Psi}(P_n)$ implies for any cross-

validation scheme B_n

$$\begin{aligned} B(P_n) &= E_{B_n}\hat{\Psi}(P^0_{nB_n}) - \hat{\Psi}(P_n) \\ &= E_{B_n}\left\{\hat{\Psi}(P^0_{nB_n}) - \hat{\Psi}(P_0)\right\} - \left\{\hat{\Psi}(P_n) - \hat{\Psi}(P_0)\right\} \\ &= E_{B_n}\left\{(P^0_{nB_n} - P_0)D(P_0) + R(P^0_{n,B_n})\right\} \\ &\quad - \left\{(P_n - P_0)D(P_0) + R(P_n)\right\} \\ &= E_{B_n}R(P^0_{n,B_n}) - R(P_n), \end{aligned}$$

where we use that $E_{B_n}P^0_{n,B_n}D(P_0) = P_nD(P_0)$. Thus, our proposed bias estimate $B(P_n)$ *equals*, for any cross-validation scheme, an average difference of the remainder applied to a subsample of size $n(1-p)$ and the full sample of size n. Therefore, one can conclude that this term will be very sensitive to a large remainder (e.g., second order terms) in the asymptotic linearity expansion (3).

6.6 MSE of targeted maximum likelihood estimator relative to its δ-limit.

If $d = 1$, then we define the MSE term as

$$MSE(P_n)(\delta) = \frac{\sigma^2(P_n)(\delta)}{n} + B(P_n)^2.$$

If $d > 1$, then we assume that we are provided with a user-specified $d \times d$ symmetric positive definite matrix ρ, so that the square root of this matrix $\rho^{1/2}$ exists. Our MSE term will represent the expectation of the Euclidean norm of $\rho^{1/2}(\hat{\Psi} - \psi)$, or equivalently, the expectation of $(\hat{\Psi} - \psi)^\top\rho(\hat{\Psi} - \psi)$. One concrete proposal is to set $\rho^{1/2}$ equal to the square root of the inverse of an estimate of the correlation matrix of the asymptotic covariance matrix of $\sqrt{n}(\hat{\Psi}-\psi)$, so that the linearly transformed vector has uncorrelated components.

Let a_j be the j-th row of the matrix $\rho^{1/2}$, $j = 1, \ldots, d$. The wished MSE term is now the sum of the MSEs of the linear combination $a_j^\top\hat{\Psi}$. Therefore, the MSE term is represented as

$$MSE(P_n)(\delta) = \frac{1}{n}\sum_j a_j^\top\Sigma(P_n)(\delta)a_j + n\left\{a_j^\top B(P_n)(\delta)\right\}.$$

This is equivalent to defining a variance term

$$\frac{\sigma^2(P_n)(\delta)}{n} = \frac{1}{n}\sum_j a_j^\top \Sigma(P_n)(\delta) a_j,$$

a bias term

$$b(P_n)(\delta) = \sum_j \left\{ a_j^\top B(P_n)(\delta) \right\},$$

and defining

$$MSE(P_n)(\delta) = \frac{\sigma^2(P_n)(\delta)}{n} + \{b(P_n)(\delta)\}^2.$$

6.7 Scaling the MSE term relative to the log-likelihood terms.

This subsection explores some issues regarding the scaling of the MSE term relative to the log-likelihood terms.

We will subtract the $MSE(P_n)(\delta)$ from the log-likelihood driven criterion $L_{CV}(P_n)(\delta) + L^*(P_n)(\delta)$ or $L^*_{CV}(P_n)(\delta)$, but we wish to do so in a way that achieves suitable balance between an increase in the log-likelihood fit and an increase in $MSE(P_n)(\delta)$. Inspection of the log-likelihood might make the choice of scaling reasonably obvious, but here we wish to present some calculations that might shed further light on this scaling issue.

For the sake of illustration, suppose $d = 1$. Note that we then view $MSE(P_n)(\delta)$ as an estimate of the squared distance $(\psi_n(\delta) - \psi_0(\delta))^2$ of the δ-specific targeted maximum likelihood estimator and its target/limit $\psi_0(\delta)$. On the other hand, we view $L_{CV}(P_n)(\delta) + L^*(P_n)(\delta)$ as an estimate of the Kullback-Leibler dissimilarity $d_{KL}(\hat{P}^*_\delta(P_n), P_0) \equiv P_0 \log d\hat{P}^*_\delta(P_n)/dP_0$ between the δ-specific targeted maximum likelihood estimator $\hat{P}^*_\delta(P_n)$ and the wished target P_0. Therefore, from that perspective, the right balance concerns the relation between the global dissimilarity $d_{KL}(P, P_0)$ and the targeted dissimilarity $(\Psi(P) - \Psi(P_0))^2$.

Our suggestion is that a meaningful quadratic distance might be inspired by $d_{KL}(P_0, P_0(\epsilon))$ along an optimal fluctuation function $P_0(\epsilon)$ that maximally changes the target parameter along ϵ.

For that purpose, let $P_0(\epsilon)$ be a fluctuation through P_0 at $\epsilon = 0$ whose score (for each ϵ_j) equals a component of an orthogonal decomposition of the (j-th component of the) efficient influence curve or the whole (j-th component

of the) efficient influence curve. For example, suppose that the efficient influence curve $D = D_1 + D_2$ and the score of $P_0(\epsilon)$ at $\epsilon = 0$ equals $D_1(P_0)$. A second order Taylor expansion of

$$d_{KL}(P_0, P_0(\epsilon)) = P_0 \log dP_0/dP_0(\epsilon)$$

at $\epsilon = 0$ shows that

$$d_{KL}(P_0, P_0(\epsilon)) \approx \epsilon^\top I(0)\epsilon,$$

where $I(0) = d/d\epsilon P_0 D_1(P_0(\epsilon))|_{\epsilon=0}$. By pathwise differentiability of Ψ we also have

$$\begin{aligned}\Psi(P_0(\epsilon)) - \Psi(P_0) &\approx P_0 D(P_0)\frac{dP_0(\epsilon) - dP_0}{dP_0} \\ &\approx P_0 D(P_0)(\sum_j \epsilon(j) D_{1j}(P_0)) \\ &= \sum_j P_0 D_1(P_0) D_{1j}(P_0).\end{aligned}$$

If we define

$$I^*(0)(k,l) = P_0 D_{1k}(P_0) D_{1l}(P_0),$$

then it follows that

$$\Psi(P_0(\epsilon)) - \Psi(P_0) \approx I^*(0)\epsilon,$$

or

$$\begin{aligned}d_{I^*(0)}(\Psi(P_0(\epsilon)), \Psi(P_0)) &\equiv (\Psi(P_0(\epsilon)) - \Psi(P_0))^\top I^*(0)^{-1}(\Psi(P_0(\epsilon)) - \Psi(P_0)) \\ &\approx \epsilon^\top [I^*(0)]\epsilon.\end{aligned}$$

Since $d_{KL}(P_0(\epsilon), P_0) \approx \epsilon^\top I(0)\epsilon$ is on the same scale as $\epsilon^\top I^*(0)\epsilon$ it follows that

$$d_{KL}(P_0, P_0(\epsilon)) + d_{I^*(0)}(\Psi(P_0(\epsilon)), \Psi(P_0))$$

is an appropriately scaled dissimilarity measure.

In the nonparametric causal effect example we have $D_1(P_0(\epsilon)) = h_0(A,W)(Y - Q_0(A,W) - \epsilon h_0(A,W))$, so that $I(0) = P_0 h_0^2$ and $I^*(0) = P_0 h_0^2 (Y - Q_0)^2$. So in that case we would have to divide the MSE by $I^*(0)$.

However, if one would use the log-likelihood under a normal error regression model and set the variance of the residuals equal to 1 so that the RSS

can be equated with the log-likelihood, then this factor can be set equal to 1.

We suggest that one should aim to scale the MSE term in such a way that it is put on equal footing with the log-likelihood gain along an optimal fluctuation function. The above calculations suggests that this can be achieved by making the MSE term representative of the standardized Euclidean dissimilarity $d_{I^*(0)}(\psi, \psi_0)$.

7 Example: Targeted maximum likelihood estimation of the marginal structural model.

Suppose we observe $O = (W, A, Y = Y(A))$, where W are baseline covariates, A is a discrete treatment, and Y is a subsequently measured outcome. It is assumed that A is realized in response to the realization of W, and Y is realized in response to both W and A. The full data structure on the experimental unit is $X = (W, (Y(a) : a))$, so that A represents the missingness variable for the missing data structure O on X.

Consider a marginal structural model for the full data distribution

$$E_0(Y(a) \mid V) = m(a, V \mid \beta_0)$$

that models the causal effect of a treatment intervention $A = a$ on the outcome Y. For example, one might assume a simple linear model $m(a, V \mid \beta_0) = \beta_0(a, V, aV)$.

Since it is often unreasonable to assume such a parametric form, but such parametric forms can still provide very meaningful projections of the true causal curve, we consider its nonparametric extensions:

$$\Psi_h(P_0) = \underset{\beta}{\operatorname{argmin}}\ E_{P_0} \sum_a h(a, V)(Q_0(a, W) - m(a, V \mid \beta))^2,$$

where $Q_0(a, w) = E_0(Y \mid A = a, W)$. If the randomization assumption that $A \perp X$, given W holds, so that, $g_0(a \mid X) = P_0(A = a \mid X) = P_0(A = a \mid W)$, then $\Psi_h(P_0)$ represents a projection of $E_0(Y(a) \mid V)$ onto the working model $m(\mid \beta_0)$. That is,

$$\Psi_h(P_0) = \underset{\beta}{\operatorname{argmin}}\ E_{P_0} \sum_a h(a, V) \left(E(Y(a) \mid V) - m(a, V \mid \beta)\right)^2.$$

In particular, if $E_0(Y(a) \mid V) = m(a, V \mid \beta_0)$, then for each h we have $\Psi_h(P_0) = \beta_0$. Without the randomization assumption, we can interpret $\Psi_h(P_0)$ as an effect of A on Y that controls for the confounders W, which we often refer to as a variable importance measure.

We note that this nonparametric extension only depends on P_0 through the conditional mean of Y, given A, W, and the marginal distribution of W. For simplicity, we will also use the notation $\Psi_h(Q_0)$, where Q_0 now denotes both the marginal distribution of W and the conditional distribution of Y, given A, W.

The efficient estimating function for this nonparametric extension Ψ_h of β_0 is given by:

$$\begin{aligned} D_h(P_0)(O) &= \frac{h_1(A,V)}{g_0(A \mid W)}(Y - Q_0(A,W)) \\ &\quad + \sum_a h_1(a,V)(Q_0(a,W) - m(a,V \mid \Psi_h(P_0)), \end{aligned}$$

where $h_1(a,V) = h(a,V)\frac{d}{d\psi}m(a,V \mid \psi)$. We will assume that $h_1(A,V) = d/d\psi_0 m(A,V \mid \psi_0)h(A,V)$ is chosen so that h_1 does not depend on ψ_0, which is easily arranged for the case that m is linear in ψ and that m is logistic linear. Let $D_h^*(P_0) = -c_0^{-1}D_h(P_0)$ be the corresponding efficient influence curve obtained by standardizing the efficient estimating function by the negative of the inverse of the derivative matrix $c_0 = d/d\psi_0 E_0 D_h(P_0)$ (noting that $D_h(P_0)$ can indeed be viewed as function in ψ_0).

If Y is continuous and we use a normal error regression model as a working model, then a targeted maximum likelihood estimator of ψ_{h0} can be obtained by adding to an initial estimator $Q^0(A,W)$ of $E_0(Y \mid A, W)$ the d-dimensional ϵ-extension $\epsilon C_h(g)(A,W)$, where

$$C_h(g)(A,W) = \frac{h_1(A,V)}{g(A \mid W)},$$

for some fit g of g_0, and fitting ϵ with maximum likelihood estimation using Q^0 as offset. The resulting update $Q^1(A,W)$ is now a first step targeted maximum likelihood estimator. One estimates the distribution of W with the empirical distribution. The estimate Q^1 and the empirical distribution of W now yields a substitution estimate of the target parameter ψ_{h0}.

Similarly, the same covariate extension can be used in a logistic regression fit of Q_0 if Y is binary.

7.1 Penalized log-likelihood for candidate treatment mechanism fits.

Let $\hat{Q}(P_n)$ be an initial regression estimator of $Q_0 = E_0(Y \mid A, W)$. For a given $P_n \to \hat{g}(P_n)$, let $\hat{Q}^*_{\hat{g}}(P_n)$ be the targeted maximum likelihood estimator corresponding with the covariate $C_h(\hat{g}(P_n))$. Let B_n be a cross-validation scheme, and let P^1_{n,B_n} and P^0_{n,B_n} be the empirical distributions of the validation and training sample, respectively, as identified by $B_n \in \{0,1\}^n$. Let

$$\hat{\Sigma}_{CV}(P_n)(\hat{g}) = E_{B_n} P^1_{n,B_n} D^*_h(\hat{Q}^*_{\hat{g}}(P^0_{n,B_n}), \hat{g}(P^0_{n,B_n}))^2$$

be the cross-validated estimate of the covariance matrix of the efficient influence curve at the estimator $\hat{Q}$ and a certain $\hat{g}$. We also consider the empirical estimate of this covariance matrix

$$\hat{\Sigma}(P_n)(\hat{g}) = P_n D^*_h(\hat{Q}^*_{\hat{g}}(P_n), \hat{g}(P_n))^2.$$

Let

$$\hat{B}(P_n)(\hat{g}) = E_{B_n} \hat{\Psi}_{\hat{g}}(P^0_{n,B_n}) - \hat{\Psi}_{\hat{g}}(P_n)$$

be the bias estimator for the targeted maximum likelihood estimator $\hat{\Psi}_{\hat{g}}(P_n) = \Psi_h(\hat{Q}^*_{\hat{g}}(P_n))$ obtained by plugging in $\hat{Q}^*_{\hat{g}}(P_n)$ in the parameter mapping Ψ.

We will penalize the log-likelihood with normally distributed residuals, or equivalently, the sum of squared residuals, with the estimate of the mean squared error

$$\frac{1}{n}\sum_{i=1}^{n}(E(m(a_i, v_i \mid \hat{\Psi}_{\hat{g}}(P_n))) - m(a_i, v_i \mid \hat{\Psi}_{\hat{g}}(P_0)))^2.$$

This mean squared error can be decomposed as $1/n \sum_{i=1}^n \mathrm{Var}(m(a_i, v_i \mid \hat{\Psi}_{\hat{g}}(P_n)))$ and $1/n \sum_{i=1}^n \mathrm{Bias}^2(m(a_i, v_i \mid \hat{\Psi}_{\hat{g}}(P_n)))$. The variance terms of this mean squared error can be estimated by

$$\frac{\sigma_i^2(\hat{g})}{n} \equiv \frac{z(a_i, v_i)^\top \hat{\Sigma}(P_n)(\hat{g}) z(a_i, v_i)}{n},$$

where

$$z(a_i, v_i) = \frac{d}{d\beta} m(a_i, v_i \mid \beta)\Big|_{\beta = \hat{\Psi}_{\hat{g}}(P_n)}.$$

We keep open the option that one uses either the cross-validated covariance matrix $\hat{\Sigma}_{CV}(P_n)$ or the empirical covariance matrix $\hat{\Sigma}(P_n)$.

The bias terms of this mean squared error can be estimated as

$$B_i(\hat{g}) \equiv E_{B_n} m(a_i, v_i \mid \hat{\Psi}_{\hat{g}}(P^0_{n,B_n})) - m(a_i, v_i \mid \hat{\Psi}_{\hat{g}}(P_n)).$$

If m is linear in β, then the latter reduces to

$$B_i(\hat{g}) = m(a_i, v_i \mid B(P_n)).$$

Thus, we obtain the following mean squared error estimate for the targeted maximum likelihood estimator $\hat{\Psi}_{\hat{g}}(P_n)$ for a given g-estimator:

$$\widehat{MSE}(P_n)(\hat{g}) \equiv \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\sigma_i^2(\hat{g})}{n} + B_i(\hat{g})^2 \right\}.$$

We suggest that the penalized log-likelihood could also only be penalized by the empirical variance component of the MSE. Therefore, we also define

$$\sigma^2(P_n)(\hat{g}) \equiv \frac{1}{n} \sum_{i=1}^{n} \frac{\sigma_i^2(\hat{g})}{n}.$$

Consider now the following two penalized log-likelihood criterions for $\hat{g}$, given the initial estimator $\hat{Q}^0$:

$$L(\hat{g} \mid \hat{Q}^0) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Q}^*_{\hat{g}}(P_n)(W_i, A_i))^2 + \widehat{MSE}(P_n)(\hat{g}),$$

or

$$L(\hat{g} \mid \hat{Q}^0) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Q}^*_{\hat{g}}(P_n)(W_i, A_i)))^2 + \sigma^2(P_n)(\hat{g}).$$

7.2 Algorithm for estimating the treatment mechanism based on penalized log-likelihood.

Given any candidate adjustment set $W^* \subset W$, let an estimator $\hat{g}(P_n)(W^*)$ of $g_0(A \mid W^*)$ be specified.

This allows us to define a criterion in adjustment sets W^*, given the current estimator $\hat{Q}$:

$$L(W^* \mid \hat{Q}) \rightarrow L(\hat{g}(P_n)(W^*) \mid \hat{Q}).$$

Given $\hat{Q}$, one can now use this empirical criterion in adjustment sets to construct an estimator of $g_0(\hat{Q})$ with a greedy type algorithm maximizing over a set of candidate adjustment sets. Firstly, one can evaluate any given adjustment set W^* with $L(W^* \mid \hat{Q})$. One starts with the empty adjustment set and selects the best addition move among a set of candidate addition moves based on the criterion. One iterates this process until there does not exist an addition move that improves the criterion. More aggressive greedy algorithms can be implemented as well, as with any machine learning algorithm that is based on iterative local maximization of an empirical criterion.

Alternatively, one creates a sequence of nested (increasing in size) adjustment sets W_j^*, $j = 1, \ldots, J$, for each W_j^* one obtains a particular estimator $\hat{g}_j(P_n)$ of $g_0(A \mid W_j^*)$ (e.g., using super learning), and maximizes the penalized log-likelihood criterion over all these J adjustment sets.

In our algorithm in the next subsection defining the sequence of C-TMLEs we apply this greedy algorithm to candidate estimators that are more nonparametric than the selected estimator of g_0 in the previous step.

7.3 Iteration to obtain sequence of collaborative targeted maximum likelihood estimators.

Given an initial estimator $\hat{Q}$ of $E(Y \mid A, W)$ and a corresponding estimator $\hat{g}(\hat{Q})$, sometimes denoted with $\hat{g}$, we define a resulting targeted maximum likelihood estimator

$$\hat{Q}^*_{\hat{g}}(P_n) = \hat{Q}(P_n) + \epsilon_n h(\hat{g}(\hat{Q})(P_n)),$$

where ϵ_n is the least squares estimator of the regression coefficient ϵ treating $\hat{Q}(P_n)$ as offset and $h(\hat{g}(\hat{Q})(P_n))$ as covariate. We can define this as a first step targeted maximum likelihood estimator based on an initial $\hat{Q}(P_n)$, and corresponding censoring mechanism estimator $\hat{g}(\hat{Q})$). Let's denote this operation as:

$$\hat{Q}^1(P_n) = \hat{Q}(P_n) + \epsilon_n^1 h(\hat{g}(\hat{Q})(P_n)).$$

This process can now be iterated by replacing $\hat{Q}(P_n)$ by this update $\hat{Q}^1(P_n)$:

$$\hat{Q}^2(P_n) = \hat{Q}^1(P_n) + \epsilon_n^2 h(\hat{g}(\hat{Q}^1)(P_n)),$$

where we require that the next censoring mechanism estimator $\hat{g}(\hat{Q}^1)(P_n)$ is obtained with the same algorithm as above, but now maximizing over candidate estimators that are more nonparametric than $\hat{g}(\hat{Q})(P_n)$.

In general, we define the k-th step of this targeted maximum likelihood estimator as

$$\hat{Q}^k(P_n) = \hat{Q}^{k-1}(P_n) + \epsilon_n^k h(\hat{g}(\hat{Q}^{k-1})(P_n)),$$

where $\hat{g}(\hat{Q}^{k-1})(P_n)$ involves maximizing over more nonparametric candidate estimators than $\hat{g}(\hat{Q}^{k-2})(P_n)$.

This algorithm results in a sequence of collaborative targeted maximum likelihood estimators $\Psi(\hat{Q}^k(P_n))$ of ψ_0, and corresponding increasingly nonparametric censoring mechanism estimators $\hat{g}^k(P_n)$ (i.e., $\hat{g}(\hat{Q}^{k-1})(P_n)$ in above notation), $k = 1, \ldots, K$.

7.4 Selection among different candidate TMLEs.

If the initial estimator $\hat{Q}$ is indexed by a choice δ_1 and the choice of algorithm $\hat{g}(\hat{Q})$ is indexed by a δ_2, then this results in candidate collaborative targeted maximum likelihood estimators $P_n \to \hat{Q}^k_{\delta_1,\delta_2}(P_n)$, corresponding treatment mechanism estimators $P_n \to \hat{g}^k_{\delta_2}(P_n)$, and corresponding $P_n \to \Psi(\hat{Q}^k_{\delta_1,\delta_2}(P_n))$ targeted maximum likelihood estimators of ψ_0, indexed by triplets (k, δ_1, δ_2).

In order to select among these candidate targeted maximum likelihood estimators indexed by (k, δ_1, δ_2) we use our proposed cross-validated penalized log-likelihood defined as

$$\begin{aligned} L(k, \delta_1, \delta_2) &= E_{B_n} P^1_{n,B_n} \left(Y - \hat{Q}^k_{\delta_1,\delta_2}(P^0_{n,B_n})(W, A) \right)^2 \\ &\quad + \widehat{MSE}_{CV}(P_n)(\hat{Q}^k_{\delta_1,\delta_2}, \hat{g}^k_{\delta_2}). \end{aligned}$$

7.5 Statistical inference.

The resulting collaborative targeted maximum likelihood estimator $Q_n = \hat{Q}^*(P_n)$ and corresponding $g_n = \hat{g}(P_n)$ solve the efficient influence curve equation $0 = P_n D^*(\Psi(Q_n), g_n, Q_n)$, so that $\psi_n = \Psi(Q_n)$ can be analyzed with our asymptotics theorem, and inference can be based on the influence curve.

8 Simulation.

In this section we first describe an implementation of the C-TMLE algorithm, then review other estimators in the literature before presenting the results of three simulations designed to offer a performance comparison across a variety of situations commonly found in the analysis of real-world data. Though each of the estimators described below is capable of providing an unbiased estimate of the parameter of interest under ideal conditions, results indicate that the C-TMLE estimator consistently performs as well or better than the others across all simulations. We end by comparing performance of the new C-TMLE estimator with the standard TMLE.

8.1 C-TMLE implementation.

The specific choices outlined below were used to run the simulations presented in this section and for the data analysis described in Section 9.

Step 1: Obtain a stage 1 estimate Q_n^0 of $Q(A, W)$. Though super learning to determine optimal weighted combinations of candidate machine learning algorithms is recommended, any particular data adaptive machine learning algorithm providing a consistent estimate is acceptable. For these simulations the DSA algorithm introduced in Sinisi and van der Laan (2004) was used to provide the initial estimate of the the true regression of Y on treatment A and confounders W.

Step 2: Generate candidate second stage estimators Q_n^k. A super learner implementation described below is recommended for this step, however in these simulations forward selection was used to build a sequence of updates for g that are increasing in size.

Though not required, a sensible approach is to use the intercept model for g to construct the covariate, h_1, used to create the first targeted maximum likelihood candidate, Q_n^1.

$$g_1(1 \mid W) = P(A = 1),\ g_1(0 \mid W) = P(A = 0)$$

$$h_1 = \left(\frac{I[A = 1]}{g_1(1 \mid W)} - \frac{I[A = 0]}{g_1(0 \mid W)} \right)$$

$Q_n^1 = Q_n^0 + \epsilon_1 h_1$, where ϵ_1 is fitted by regressing Y on h with offset Q_n^0. Next we create an updated model for g. The intercept term is forced

into the next model for g. Additional terms are incorporated as long as they increase the overall penalized likelihood. Increasing the penalized likelihood is equivalent to decreasing the sum of the empirical sum of squared residuals plus the empirical variance of the efficient influence curve (see below) at the updated Q-fit and the candidate g-fit. In the event that no terms increase the penalized likelihood, the term that provides the best updated penalized likelihood is forced into the model.

As an example suppose that in addition to the intercept term, m terms, ordered $1, \ldots, m$, are incorporated into the model, at which point no further increase of the penalized likelihood is possible. We define candidate estimators Q_n^2 through Q_n^{m+1} as:

$$\begin{aligned} Q_n^2 &= Q_n^1 + \epsilon_2 h_2 \\ Q_n^3 &= Q_n^1 + \epsilon_3 h_3 \\ &\vdots \\ Q_n^{m+1} &= Q_n^1 + \epsilon_{m+1} h_{m+1} \end{aligned}$$

where the model for g_n^{i+1} contains all the terms in the model for g_n^i plus one additional term. At this point Q_n^{m+1} is considered as a new "initial" estimate of the true regression, and the entire process starts over in order to build a second clever covariate augmenting the previous fit g_n^{m+1} used in h_{m+1}. To continue the example, $Q_n^{m+2} = Q_n^{m+1} + \epsilon_{m+2} h_{m+2}$. This process is iterated until all terms are incorporated into the final model for g. If the maximal number of terms that can be added is given by K, then this results in K candidate estimators Q_n^k, $k = 1, \ldots, K$, corresponding with treatment mechanism estimators g_n^k, $k = 1, \ldots, K$. Note that the number of clever covariates in Q_n^k that are added to the initial estimator Q_n^0 cannot be predicted, and depends on how many covariates can be added to the treatment mechanism estimator in each iteration before reaching the local maximum (not allowing a further increase of the penalized log-likelihood).

Note that the model for g is not restricted to main terms only. For example, variables can be created that correspond to higher-order terms. In addition, a categorical or continuous covariate can be split into many binary covariates, thereby allowing for more nonparametric modelling of the effect of a single covariate. When there are many covariates it

might be desirable in practice to terminate the procedure before all covariates have been incorporated into the model for g, though care must be taken to ensure that none of the candidates thereby excluded from the subsequent selection process potentially maximize the penalized log-likelihood criterion.

A superior version of this procedure uses the super learner to estimate the series of increasingly nonparametric candidate TMLEs. For this version we use the forward selection algorithm just described to obtain an ordering $1, \ldots, K$ over all potential confounders in the set W. The super learner uses this ordering to estimate each treatment mechanism when $k = 1, 2, \ldots, K$ that is superior to an estimate based on a main-terms regression alone, except on those rare occasions when the true treatment mechanism is a function of main terms only. (Of course, a main-terms regression model can be incorporated into the super learner, as well.) As in the algorithm presented above, the number of clever covariates used to update the initial estimator Q_n^0 depends entirely on the likelihood and cannot be pre-determined. Terms are incorporated into the model for g for a single clever covariate until there is a decrease in the likelihood. At that point the density estimate is updated from $Q_n^m \rightarrow Q_n^{(m+1)}$ and the process re-iterates until all candidate TMLEs have been constructed.

We also note that we can represent these estimators Q_n^k and corresponding treatment mechanism estimators g_n^k as mappings $\hat{Q}^k$ and $\hat{g}^k$ applied to the empirical distribution P_n: $Q_n^k = \hat{Q}^k(P_n)$, $g_n^k = \hat{g}^k(P_n)$, $k = 1, \ldots, K$. These mappings $P_n \rightarrow \hat{Q}^k(P_n)$ represent our candidate estimators of the true regression Q_0, and in the next step we use cross-validation to select among these candidate algorithms.

Step 3: Select the estimator that maximizes the V-fold cross-validated penalized likelihood, where V was set to 5. Maximizing the penalized likelihood is equivalent to minimizing the residual sum of squares (RSS) plus a penalty term corresponding to the mean squared error (MSE), which can be decomposed into variance and bias terms:

$$k^* = \underset{k}{\operatorname{argmin}}\ cvRSS_k + cvVar_k + n * cvBias_k^2.$$

These terms are defined as follows:

$$
\begin{aligned}
cvRSS_k &= \sum_{v=1}^{V} \sum_{i \in Val(v)} (Y_i - \hat{Q}^k(P_{nv}^0)(W_i, A_i))^2 \\
cvVar_k &= \frac{1}{V} \sum_{v=1}^{V} \text{var}(IC_v(\hat{Q}^k(P_{nv}^0), g_n^k)) \\
cvBias_k &= \frac{1}{V} \sum_{v=1}^{V} \Psi(\hat{Q}^k(P_n)) - \Psi(\hat{Q}^k(P_{nv}^0)) \\
IC_v(Q, g) &= \sum_{i \in Val(v)} \frac{I[A_i = 1] - I[A_i = 0]}{g(A_i \mid W_i)} (Y_i - Q(A_i, W_i)) \\
&\quad + \sum_{i \in Val(v)} Q(W_i, 1) - Q(W_i, 0) - \Psi(Q),
\end{aligned}
$$

where v ranging from 1 to V indexes the validation set $Val(v)$ for the vth fold, and $\hat{Q}^k(P_{nv}^0)$ denotes the k-th C-TMLE applied to the corresponding training sample P_{nv}^0.

A standard error and 95% confidence interval for the C-TMLE estimator can be constructed based on the variance of the efficient influence curve (IC): $\text{SE}(\psi_n) = \sqrt{var(IC)/n}$. A 95% confidence interval is given as $\psi_n \pm 1.96\text{SE}(\psi_n)$. Diagnostic tests not reported in this paper indicate that confidence intervals constructed in this way achieve the desired coverage rate, but this result is not theoretically grounded. We recommend the bootstrap procedure for inference. Alternatively, inference can be based off of a corrected influence curve for this estimator that does not ignore the contribution from Q in the case where the selected $\hat{g}$ is not a consistent estimator for g^0.

There are other methods for obtaining ψ_n^{C-TMLE}. For example, given set of candidate nuisance parameter estimators that includes highly nonparametric candidates we could use forward selection to build models of increasing size, where each term in the model corresponds to a candidate nuisance parameter estimator. The best model can be selected using likelihood-based cross-validation. Note that coefficients in front of each term are estimated by least squares, thereby solving the efficient influence equation corresponding to each nuisance parameter estimator, in particular the most nonparametric of these.

However, correlations among candidate estimators included as terms in the model are likely to result in highly variable coefficient estimates, and therefore increased variance in the estimate of the parameter of interest. Removing all but the most nonparametric candidate estimator from the selected model is an ad hoc bias/variance tradeoff not specifically targeted to the parameter of interest that did not improve performance in simulation studies. This approach is not recommended when the candidate estimator ordering can be determined.

8.2 Current methods for estimating marginal treatment effects.

We review current methods for estimating the marginal effect of a treatment A on outcome Y as a prelude to comparing performance on simulated data. The estimators under consideration in addition to C-TMLE are the G-computation estimator Robins (1986), the $IPTW$ estimator (Hernan et al. (2000), Robins (2000b)), the double robust $IPTW$ estimator ($DR-IPTW$), (Robins and Rotnitzky (2001); Robins et al. (2000); Robins (2000a)), and an extension to propensity score matching implemented in a publicly available R package (Sekhon (2008)).

Given observations $O = (W, A, Y)$, we are interested in estimating an adjusted marginal effect of treatment A on outcome Y given a vector W of potential confounders. If we restrict our discussion to the case where A is binary our parameter of interest is given by: $\psi = E_W[E[Y \mid A = 1, W] - E[Y \mid A = 0, W]]$. Each of the four estimators we are considering rely on estimates of one or both of the following nuisance parameters: $Q(A, W) \equiv E[Y \mid A, W]$ and $g(A, W) \equiv P(A \mid W)$. The first nuisance parameters can be estimated by, for example, a regression of Y on A and W. When the second nuisance parameter is unknown it can be estimated by a logistic regression of A on W. Each estimator is defined below.

$$
\begin{aligned}
\psi_n^{Gcomp} &= \frac{1}{n}\sum_{i=1}^{n} Q_n(1,W_i) - Q_n(0,W_i) \\
\psi_n^{IPTW} &= \frac{1}{n}\sum_{i=1}^{n} \left[I(A_i=1) - I(A_i=0)\right] \frac{Y_i}{g_n(A_i,W_i)} \\
\psi_n^{DR-IPTW} &= \frac{1}{n}\sum_{i=1}^{n} \frac{[I(A_i=1) - I(A_i=0)]}{g_n(A_i \mid W_i)}(Y_i - Q_n^0(W_i,A_i)) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} Q_n^0(1,W_i) - Q_n^0(0,W_i) \\
\psi_n^{C-TMLE} &= \frac{1}{n}\sum_{i=1}^{n} \frac{[I(A_i=1) - I(A_i=0)]}{g_n^*(A_i \mid W_i)}(Y_i - Q_n^*(W_i,A_i)) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} Q_n^*(1,W_i) - Q_n^*(0,W_i) \\
\psi_n^{PropScore} &= \frac{1}{m}\sum_{i=1}^{m} \left[I(A_i=1) - I(A_i=0)\right] Y_i
\end{aligned}
$$

where Q_n^0 refers to an initial untargeted estimate of $Q(A \mid W)$, Q_n^* refers to an updated targeted estimate of $Q(A \mid W)$ described in detail in the next section, and m in the last equation indexes observations matched on propensity score and covariates W.

The G-computation estimator relies on consistent estimation of Q. The $IPTW$ estimator depends on consistent estimation of g. $DR-IPTW$ yields valid estimates if one or both nuisance parameters are estimated consistently. The augmented propensity score estimator included here uses the genetic algorithm (Holland and Reitman (1977)) to determine a combination of propensity score and covariate values that provides the best matches between observations where $A = 1$ and observations where $A = 0$. The marginal treatment effect is estimated as the average over all matches. ETA violations reduce the quality of the match and introduce bias into the estimate. This method is especially effective when overall match quality is a function of true confounders. Estimates can suffer even when match quality is high if a small subset of covariates that are large confounders are not well-matched.

8.3 Comparison of estimators.

For each simulation we have a data structure $O = (W, A, Y)$, where $W = (W_1, \ldots, W_6)$ is a set of potential confounders of the relationship between binary treatment variable A and continuous outcome Y. Our parameter of interest is the marginal effect of treatment on the outcome: $\psi = E_W[E[Y \mid A = 1, W] - E[Y \mid A = 0, W]]$. The simulations are designed to demonstrate estimator performance in the face of confounding of the relationship between treatment and outcome, complex underlying data-generating distributions (e.g., high level interactions and non-linear functional forms), and practical violations of the Experimental Treatment Assumption (ETA), i.e., $P(A = a \mid W) < \alpha$, for some small α, implying that there is very little possibility of observing both treated and untreated subjects for some combination of covariates present in the data.

8.3.1 Data generation.

Covariates $W_1, \ldots W_5$ were generated as independent normal random variables. W_6 is a binary variable.

$$\begin{aligned} W_1, W_2, W_3, W_4, W_5 &\sim N(0,1) \\ logit(W_6) &= .3W_1 + .2W_2 - 3W_3 \end{aligned}$$

Two treatment mechanisms were defined:

$$\begin{aligned} logit\left(g_{1,0}(A \mid W)\right) &= .3W_1 + .2W_2 - 3W_3 \\ logit\left(g_{2,0}(A \mid W)\right) &= .15(.3W_1 + .2W_2 - 3W_3) \end{aligned}$$

The observed outcome Y was generated as

$$Y = Q_{i,0}(A, W) + \epsilon,\ \epsilon \sim N(0,1)$$

with corresponding regression equations:

$$\begin{aligned} Q_{1,0}(A, W) &= A + .5W_1 - 8W_2 + W_3 + 8W_3 - 2W_5 \\ Q_{2,0}(A, W) &= A + .5W_1 - 8W_2 + W_3 + 8W_3^2 - 2W_5 \end{aligned}$$

We consider three different data-generating distributions, $(Q_{1,0}, g_{1,0})$ in simulation 1, $(Q_{2,0}, g_{1,0})$ in simulation 2, and $(Q_{2,0}, g_{2,0})$ in simulation 3. Note that

W_6 is strongly correlated with treatment mechanism A in simulations 1 and 2 (corr=0.54), but is not an actual confounder of the relationship between A and Y. W_1, W_2, and W_3 are confounders. The linear nature of the confounding due to W_3 in simulation 1 differs from that in simulations 2 and 3, where the true functional form is quadratic. In this way simulations 2 and 3 closely mimic realistic data analysis settings in which the unknown underlying functional form is seldom entirely captured by the regression model used in the analysis. Finally, the treatment mechanism in simulations 1 and 2 leads to ETA violations ($p(A = a \mid W)$ ranges between 9×10^{-7} and 0.9999978). In simulation 3 there are no ETA violations ($0.11 < p(A = a \mid W) < 0.88$). In each simulation the true value of the parameter of interest is 1.

8.3.2 Simulation.

1000 samples of size $n = 1000$ were drawn from each data generating distribution. Marginal treatment effect estimates were calculated based on the unadjusted regression of Y on A, G-comp, $IPTW, DR-IPTW$, propensity score and C-TMLE methods. A main-effects model for G-comp and $DR-IPTW, \hat{Q}$, was obtained using the DSA algorithm with the maximum model size set to seven. The propensity score function was run using default settings except population size for each generation was increased to 200. A model for the treatment mechanism $\hat{g}$ used in $IPTW, DR-IPTW$ and propensity score estimation was also selected by DSA, again restricted to main terms. In contrast, the C-TMLE algorithm includes an aggressive search through a larger space of models to obtain an initial estimate of the density. As a proxy for the super-learner algorithm we used the DSA algorithm to select a model for $\hat{Q}$ containing at most six terms, allowing quadratic terms and two-way interactions.

8.3.3 Results.

Mean estimates of the treatment effect and standard errors are shown in Table 1 for each simulation. Estimates and 95% confidence intervals are plotted in figures 2 and 3.

Table 1: Mean estimate and standard errors for each estimator based on 1000 iterations with sample size $n = 1000$. $\psi_0 = 1$.

	Simulation 1		**Simulation 2**		**Simulation 3**	
	$\overline{\psi}_n$	SE	$\overline{\psi}_n$	SE	$\overline{\psi}_n$	SE
Unadj	−11.97	0.64	−0.98	0.91	0.29	0.86
G-comp	0.99	0.09	0.76	1.22	0.95	0.68
IPTW	−4.36	0.72	0.03	0.76	0.83	0.90
DR-IPTW	0.99	0.09	0.94	0.62	1.03	0.80
C-TMLE	0.99	0.09	1.00	0.10	1.00	0.07
PropScore	−1.22	0.82	0.54	0.73	0.96	0.25

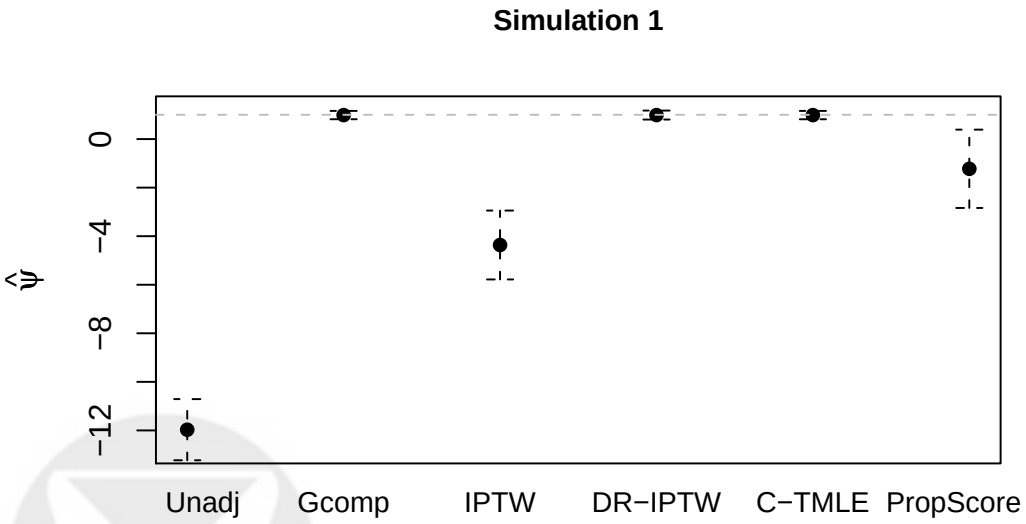
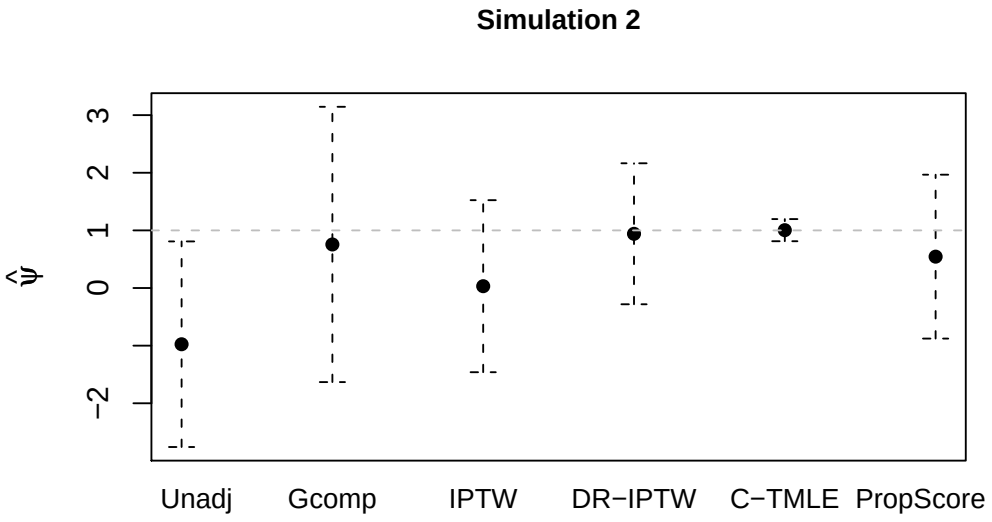


Figure 2: Estimates and 95% confidence intervals for each estimation method, simulation 1. Horizontal dashed line is at true parameter value.



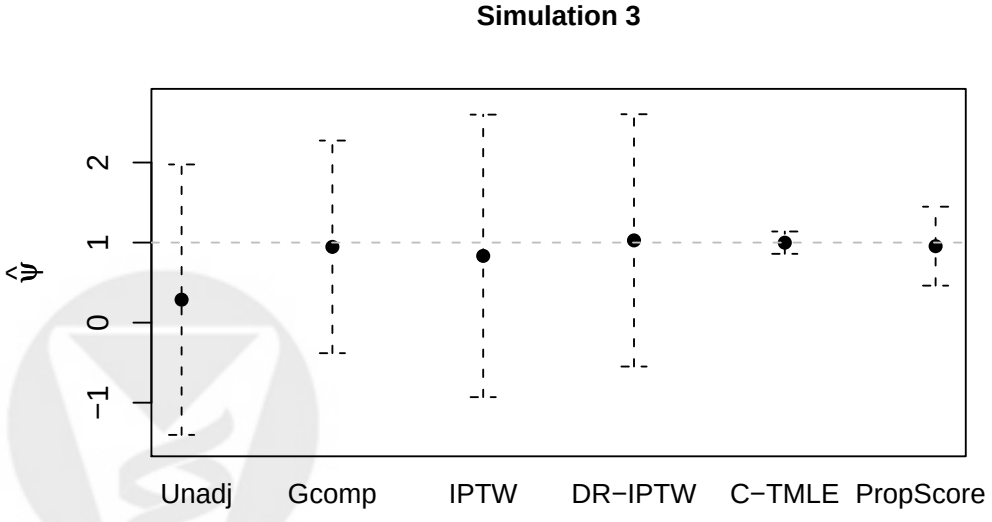


Figure 3: Estimates and 95% confidence intervals for each estimation method, simulations 2 and 3. Dashed line is at true parameter value.

Figures 2 and 3 illustrate each estimator's behavior. As expected, estimators relying on consistent estimation of Q are unbiased in simulation 1, estimators relying on consistent estimation of g are unbiased in simulation 3.

- The unadjusted estimator yields biased results in all three simulations due to its failure to adjust for confounders.
- The G-computation estimator performs well in simulation 1 when the model is correctly specified. We understand that mis-specification (simulations 2 and 3) will often, though not always, lead to bias in the estimates. However the plots highlight another phenomenon that is easy to overlook. Even when bias is not large, the inability of the mis-specified model to adequately account for the variance in the outcome often leads to large residual variance of the estimator, resulting in a failure to reject the null hypothesis.
- Truncation bias due to ETA violations causes the $IPTW$ estimator using truncated weights to fail in simulations 1 and 2. The estimate is not biased in simulation 3, but the variance is so large that even in this setting where we'd expect $IPTW$ to be reliable it fails to produce a significant result.
- $DR-IPTW$ estimates are unbiased and have low variance when the functional form is correctly modeled by the regression equation (simulation 1). Though we see little bias in the other two simulations, the variance is large due to mis-specification of the treatment mechanism. Because W_6 is a strong predictor of A and is indistinguishable from a true confounder of the relationship between Y and A it is always included in the treatment mechanism, behavior that does not help achieve an accurate estimate of the true treatment effect.
- The propensity score estimator is known to perform poorly when there are ETA violations, e.g. simulations 1 and 2 (Sekhon (2008)). It does a reasonable job in simulation 3, though the confidence interval is not as tight as for the collaborative targeted maximum likelihood estimator.
- By carefully constructing a first stage estimator of the initial density and then building a model for the treatment mechanism that adjusts

only for confounding that has not been addressed in stage 1 the C-TMLE estimator provides unbiased results with the smallest variance in all three simulated scenarios.

8.4 Comparison of C-TMLE and TMLE.

The double robust property of the targeted maximum likelihood estimator obviates the need for accurate estimation of both Q and g since correct specification of either one leads to consistent estimates of the parameter of interest. However, accurate estimates of both are needed to achieve minimum variance. Implementations of the standard targeted maximum likelihood estimator (TMLE) therefore strive for ideal estimates of both Q and g. In contrast, the collaborative nature of the second stage of the C-TMLE estimation algorithm leads to selection of an estimator for g that includes only that portion of the treatment mechanism needed to reduce bias not already adequately addressed by the first stage estimator for Q. In general, covariates included in the model for Q tend to not be incorporated into the model for g because they do not increase the penalized log-likelihood. At the same time, confounders that are not adequately adjusted for in the initial density estimate are quickly added to model for g unless the gain in bias reduction is offset by too great an increase in variance. When the initial estimate of the density is a very good fit for the true underlying density, TMLE and C-TMLE have similar performance. When the initial fit is less good, C-TMLE makes judicious choices regarding inclusion of covariates in the treatment mechanism, leading to lower variance. This is especially true when there are ETA violations. Data were simulated to illustrate this phenomenon.

8.4.1 Data generation.

Covariates $W_1, W2$, and W_3 were generated as independent random uniform variables over the interval $[0, 1]$. W_4 and W_5 are independent normally distributed random variables.

$$\begin{aligned} W_1, W_2, W_3 &\sim U(0,1) \\ W_4, W_5 &\sim N(0,1) \end{aligned}$$

Treatment mechanism g_0 was designed so that W_3 is highly predictive of treatment:

$$logit\left(g_0(A \mid W)\right) = 2W_1 + W_2 - 5W_3 + W_5$$

The observed outcome Y was generated as

$$Y = Q_0(A, W) + \epsilon,\ \epsilon \sim N(0, 1)$$

with corresponding regression equation:

$$Q_0(A, W) = A + 4W_1 - 5W_2 + 5W_4W_5$$

8.4.2 Simulation.

C-TMLE and TMLE estimates of the parameter of interest, again defined as $\psi = E_W[E[Y \mid A = 1, W] - E[Y \mid A = 0, W]]$, were obtained for 1000 samples of size $n = 1000$ drawn from data generating distribution (Q_0, g_0). For this study we deliberately select a mis-specified main-terms only model for Q by running the DSA algorithm on 100,000 observations drawn from that same distribution. $P(A = a \mid W)$ for these observations ranges from 0.004 to 0.996. Approximately 17% of the observations have covariates indicating that the probability of receiving treatment is less than 0.05, indicating that practical ETA violations in finite samples will cause unstable TMLE estimates.

For each iteration an initial estimate of the density, Q_n^0, was obtained by fitting the selected model, $Y = A + W_1 + W_2$, on n observations in the sample. We expect that any estimate of ψ based solely on this model is likely to be incorrect because the model fails to take into account the effect on the outcome of the missing interaction term, and also fails to adjust for the confounding effect of W_5. The targeting step for both targeted maximum likelihood estimators reduces this bias.

In order to construct the covariate used to target the parameter of interest in the updating step of the TMLE algorithm we obtain an estimate g_n of g_0 by running the DSA algorithm, allowing quadratic terms and two-way interaction terms to enter the model. This model was not fixed over the 1000 iterations; the model selection process was carried out each time a sample was drawn from the population. Similarly, covariates that were candidates for inclusion in the model for g in the second stage of the C-TMLE estimation algorithm include $(W1, \ldots, W_5,\ W_1^2, \ldots, W_5^2)$, and all two-way interaction terms (W_iW_j), where $i \neq j$.

8.4.3 Results

Results of the simulation are shown in Table 2. A small number of TMLE estimates were major contributers to the variance of that estimator. The

three highest TMLE estimates of the treatment effect were (771.914, 37.219, 9.518). It is likely that these high values arise from atypical samples containing observations that presented unusually strong ETA issues. In contrast, all C-TMLE estimates calculated from those same samples range between 0.307 and 1.698. Both estimator's average treatment effect estimates are not far from the true value, $\psi_0 = 1$, though C-TMLE. As expected, the variance of the TMLE estimator is many times larger than that of the C-TMLE estimator.

Not surprisingly, W_3, the strong predictor of treatment that is not a true confounder of the relationship between treatment and outcome, is included in every one of the 1000 models for g selected by the DSA algorithm, but is in only 35 of the models constructed in the second stage of the C-TMLE algorithm. At the same time, the interaction term W_4W_5 is included in only two out of 1000 models for g_0 selected by DSA, but is present in 576, more than half, of the collaborative models.

Table 2: Comparison of C-TMLE and TMLE estimators at different levels of truncation. Mean estimate and variance based on 1000 iterations.

	truncation level	# obs truncated	$\overline{\psi}_n$	variance
C-TMLE	∞	0	0.982	0.041
TMLE	∞	0	1.730	597.518
	40	1	1.358	162.379
	10	2	0.941	1.993
	5	9	0.915	1.680

8.5 Confidence Intervals

The variance of the uncorrected influence curve provides the basis for calculation of a 95% confidence interval for the C-TMLE estimate.

$$95\%CI = \psi^{C-TMLE} \pm 1.96\sqrt{(var(IC)/n)}$$

A confidence interval was constructed for each of the 1000 iterations in simulation 4, with Q mis-specified by a main-terms only regression model. Confidence intervals were also created for an additional 1000 samples from the same data generating distribution that were analyzed using a correct model for Q. When Q is correctly specified 93% of the confidence intervals constructed at a nominal 95% level contained the true parameter value. When Q was mis-specified confidence intervals were conservatively estimated, with 99% containing the true value.

9 Data Analysis.

We apply the C-TMLE estimator to an observational dataset previously analyzed by Bembom et al. (2008) and Bembom et al. (2007) with the goal of identifying HIV mutations that affect response to the antiretroviral drug lopinavir. The data includes observations, $O = (W, A, Y)$, where the outcome, Y, is the change in $\log_{10}$ viral load measured at baseline and at follow-up after treatment has been initiated. If follow-up viral load was beneath the limit of detection Y was set to the maximal change seen in the population. $A \in \{0, 1\}$ is an indicator of the presence or absence of a mutation of interest, taking on the appropriate value for each of the 26 candidate mutations in 26 separate analyses. W consists of 51 covariates including treatment history, baseline characteristics, and indicators of the presence of additional HIV mutations. Practical ETA violations stemming from high correlations among some of the covariates and/or low probability of observing a given mutation of interest make it difficult to obtain stable low variance estimates of the association between A and Y. Bembom used a targeted maximum likelihood estimation approach incorporating data-adaptive selection of an adjustment set that relies on setting a limit on the maximum allowable truncation bias introduced by truncating treatment probabilities less than α to some specified lower limit. Covariates whose inclusion in the adjustment set introduces an unacceptable amount of bias are not selected. That study's findings showed good greement with Stanford HIVdb mutation scores, values on a scale of 0 to 20 (http://hivdb.stanford.edu, as of September, 2007, subsequently modified), where 20 indicates evidence exists that the mutation strongly inhibits response to drug treatment and 0 signifies that the mutation confers no resistance. Because the C-TMLE method includes covariates in the treatment mechanism only if they improve the targeting of the parameter

of interest without having too adverse an effect on the MSE, we expect similar performance without having to specify truncation levels or an acceptable maximum amount of bias.

9.1 Analysis description.

The dataset consists of 401 observations on 372 subjects. Correlations due to the few subjects who contributed more than one observation were ignored. Separate analyses was carried out for each mutation. In each, an initial density estimate, Q_n^0, was obtained using DSA restricted to addition moves only to select a main-terms only model containing at most 20 terms, where candidate terms in W include pre-computed interactions detailed in Bembom et al. A was forced into the model. An estimate of the effect on change in viral load was recorded for each mutation. The variance of the estimate was calculated from the variance of the influence curve evaluated on the observations in the dataset and used to calculate confidence intervals.

9.2 Results.

Table 3 lists the Stanford mutation score associated with each of the HIV mutations under consideration, as well as the C-TMLE estimate of the adjusted effect of mutation on lopinavir resistance. The variance of the uncorrected influence function was used to calculate 95% confidence intervals. Confidence intervals entirely above zero indicate a mutation increases resistance to lopinavir. Eight of the twelve mutations having a mutation score of 10 or greater fall into this category. Point estimates for the remaining four mutations were positive, but the variance was too large to produce a significant result. Only one of the six mutations thought to confer slight resistance to lopinavir was flagged by the procedure, though with the exception of p10FIRVY point estimates were positive. Stanford mutation scores of 0 for four of the five mutations found to have a significantly negative effect on drug resistance support the conclusion that these mutations do not increase resistance, but are not designed to offer confirmation that a mutation can decrease drug resistance. However, Bembom et al. report that there is some clinical evidence that two of these mutations, 30N and 88S, do indeed decrease lopinavir resistance.

Our findings are quite consistent with the Stanford muation scores and with the results from the previous analysis using the data-adaptively selected

adjustment set targeted maximum likelihood estimation approach. The C-TMLE method was able to achieve these results without relying on ad hoc or user-specified tuning parameters.

Table 3: Stanford score (2007), C-TMLE estimate and 95% confidence interval for each mutation. Starred confidence intervals do not include 0.

mutation	score	estimate	95% CI
p50V	20	1.703	(0.760, 2.645)*
p82AFST	20	0.389	(0.084, 0.695)*
p54VA	11	0.505	(0.241, 0.770)*
p54LMST	11	0.369	(0.002, 0.735)*
p84AV	11	0.099	(-0.130, 0.329)
p46ILV	11	0.046	(-0.222, 0.315)
p48VM	10	0.306	(-0.162, 0.774)
p47V	10	0.805	(0.282, 1.328)*
p32I	10	0.544	(0.312, 0.777)*
p90M	10	0.209	(-0.058, 0.476)
p82MLC	10	1.610	(1.330, 1.890)*
p84C	10	0.602	(0.471, 0.734)*
p33F	5	0.300	(-0.070, 0.669)
p53LY	3	0.214	(-0.266, 0.695)
p73CSTA	2	0.635	(0.278, 0.992)*
p24IF	2	0.229	(-0.215, 0.674)
p10FIRVY	2	−0.266	(-0.522,-0.011)*
p71TVI	2	0.019	(-0.243, 0.281)
p30N	0	−0.440	(-0.853,-0.028)*
p88S	0	−0.474	(-0.840,-0.108)*
p88DTG	0	−0.426	(-0.842,-0.010)*
p36ILVTA	0	0.272	(-0.001, 0.544)
p20IMRTVL	0	0.178	(-0.111, 0.467)
p23I	0	0.822	(-0.050, 1.694)
p16E	0	0.239	(-0.156, 0.633)
p63P	0	−0.131	(-0.392, 0.131)

10 Discussion.

For most data sets little to no knowledge is available about the data generating distribution. As a consequence, the true model is a large infinite dimensional semi-parametric model. In such models there are many data adaptive approaches that can be considered for fitting the true distribution of the data, based on different approximation function spaces, different searching strategies for maximizing an empirical criterion (such as the empirical log-likelihood) over these spaces, and different methods for selecting the fine tuning parameters indexing the function spaces and search strategies. Each of these algorithms operates within the semi-parametric model, so that none of them should have any preference a priori. However, depending on the true data generating distribution, these algorithms will have very different levels of performance in approximating the true data generating distribution. As a consequence, cross-validation based super learning should be employed to find the best weighted combination among a large user supplied set of candidate estimators of the true data generating distribution. The oracle property of the cross-validation selector (van der Vaart et al. (2006), van der Laan et al. (2006)) teach us that the super learner will asymptotically perform exactly as well, w.r.t. the Kullback-Leibler dissimilarity measure, as the best weighted combination of the candidate algorithms optimized for each data set.

Even though the super learning application is an important advance over relying on any one particular estimator, it represents a best fit for the purpose of estimation of the whole distribution of the data, so that the bias-variance trade-off is not targeted w.r.t. the parameter of interest.

Therefore, our methodology involves a second targeted modification of the first stage super learner fit that aims to reduce the bias w.r.t the target parameter, while simultaneously increasing the likelihood fit. This is achieved by first determining the single fluctuation function that would yield asymptotic optimal bias reduction as defined by the efficient influence curve of the target parameter. This fluctuation function needs to have a score-vector at zero fluctuation whose linear span includes the efficient influence curve of the target parameter. This fluctuation function depends on an unknown nuisance parameter of the data generating distribution, such as a censoring mechanism.

We now define an iterative sequence of subsequent fluctuations, starting with the initial super learner fit, where the subsequent fluctuation functions are estimated with increasingly nonparametric estimates of the nuisance

parameter, including a final fully non-parametrically estimated fluctuation function. In addition, by construction, we make sure that for each fluctuation function the nuisance parameter estimator that results in maximal increase in likelihood fit is selected, among the candidate nuisance parameter estimators that are more nonparametric than the one selected at previous fluctuation function. In this way, we arrange that most of the targeted bias reduction occurs in the first few fluctuations. The actual number of times we carry out the subsequent update is selected with likelihood based cross-validation.

Essentially, we try to move towards the asymptotically optimal bias reduction along a sequence of targeted bias reduction steps, but we stop moving towards this asymptotically optimal bias reduction when it results in a loss of likelihood fit as measured by the cross-validated log-likelihood. We also propose a finer sequence of nested targeted bias reduction steps (i.e., a finer sequence of candidate second stage estimators) whose fits contain this set of candidate-fits as a subsequence, thereby potentially providing an additional improvement in practical performance of the resulting C-TMLE.

Theoretical results teach us that this push towards the asymptotically optimal bias reduction also takes into account how well the initial estimator already approximates the true distribution, by giving preference to targeted bias reduction steps that improve the log-likelihood fit. As a consequence, the C-TMLE is able to avoid selecting irrelevant or harmful (w.r.t. relevant factor of density) fits of the nuisance parameter, even though such fits might improve the overall fit of the nuisance parameter. That is, the fit of the nuisance parameter is targeted towards our primary goal, the parameter of interest.

In addition, we propose to replace the log-likelihood in this estimator by a penalized log-likelihood, where the penalty is scaled appropriately, has negligible contribution for nicely identifiable target parameters, but blows up for fits that result in extremely variable or biased estimators of the parameter of interest. Even though the penalty's effect on the Kullback-Leibler dissimilarity is asymptotically negligible for identifiable parameters, for parameters that are borderline identifiable, this penalty can yield dramatic additional finite sample improvements to the C-TMLE. In essence, it builds in a sensible robustness of the resulting C-TMLE as an estimator of the target parameter. The penalized log-likelihood can now be used to both build candidate nuisance parameter estimators, and to select among the nested sequence of candidate second stage estimators, thereby providing additional improvements to the candidate nuisance parameter estimators (avoiding candidates

that result in bad or even horrific estimators of target parameter) as well as to the selector among them.

To summarize, this article provides a template for building likelihood based estimators in semiparametric models (i.e., machine learning algorithms) that are targeted towards a particular target feature of the distribution of the data. The combination of 1) likelihood based cross-validation to select among a variety of choices, 2) super learning to select combinations of candidate choices, 3) building a nested sequence of candidate nuisance parameter estimators that are increasingly nonparametric and a corresponding sequence of two-stage targeted maximum likelihood estimators 4) penalized log-likelihood using an appropriately scaled penalty that is responsive to the MSE of the fit for the target parameter, provides a blueprint for construction of very powerful and robust machines for estimating target parameters in all kinds of semiparametric and nonparametric models.

11 Extensions.

11.1 Super Learning for Estimation of Censoring Mechanism.

In one particular embodiment of our proposed template for collaborative TMLE we first generate a sequence of highly data adaptive estimators of the censoring mechanism based on adjustment sets that are increasing in size. For each adjustment set we propose to use a super learner as estimator of the censoring mechanism. These nested adjustment sets could be extracted from an ordered list of the covariates. A variety of orderings could be considered. One interesting proposed ordering is obtained by first running the forward main term regression approach for building a censoring mechanism based on the log-likelihood of the corresponding targeted maximum likelihood estimator, as presented in detail in our data analysis section, and then use the ordering in which the covariates entered as the ordering.

Given the sequence of super learners corresponding with adjustment sets that are increasing in size, we run our collaborative TMLE algorithm that makes sure that, given the previous clever covariate uses the super learner for the top K covariates in the list, the next clever covariate only considers censoring mechanisms for adjustment sets defined by top L covariates with $L > K$.

11.2 Factorization of the Censoring Mechanism.

Suppose now that the likelihood of the censoring mechanism factors in two or more terms, such as a treatment mechanism and right-censoring mechanism. For example, a treatment mechanism of a time-dependent treatment across multiple time points factors as $g(\bar{A} \mid X) = \prod_{j=0}^{K} g(A(j) \mid \bar{A}(j-1), X)$.

Our collaborative TMLE algorithm requires us to specify the set of candidate censoring mechanism estimators, given that the previous clever covariate selected a particular censoring mechanism estimator. We need to make sure that these candidates are more nonparametric than the previously selected one, so that for increasing number of clever covariates we will converge to an (maximally) unbiased estimator of the censoring mechanism, even though for our particular data set we might only select few clever covariates.

Suppose we obtain an ordered list of candidate increasingly nonparametric estimators for each factor of the censoring mechanism. So for two factors, this results in a matrix of candidate estimators of the censoring mechanism, and any non-decrease in both coordinates results in a more nonparametric estimator. Given the previously selected estimator, corresponding with one location in this matrix, we can consider the right-upper quadrant as the set of candidate censoring mechanism estimators to consider.

In order to obtain sensible orderings of increasingly nonparametric estimators for each factor of the censoring mechanism we might first run the forward main term algorithm, mentioned in above subsection and presented in detail in the data analysis section. This algorithm now involves selecting the next best main term to add for each factor, and also selecting between these factor-specific best moves. We can now use the ordering in which the main terms are added into a factor as an ordered list of adjustments sets for that factor of the censoring mechanism, and corresponding candidate estimators (e.g. super learners).

11.3 Interpretability of the Parameter estimate in the presence of sparse data bias.

For the sake of illustration, consider the collaborative targeted maximum penalized likelihood estimator of a treatment effect $EY(1) - EY(0)$ based on the data structure $(W, A, Y = Y(A))$. Due to the penalization of the log-likelihood certain strong confounders of the treatment effect might have been excluded in the selected adjustment set for the treatment mechanism

estimator. This raises the question if we can still interpret the estimator as an estimator of the fully adjusted treatment effect.

In order to provide a more careful interpretation of the effect estimate and its standard error, we propose the following approach.

Firstly, we report the collaborative targeted maximum penalized likelihood estimate (C-TMPLE) of the treatment effect, the treatment mechanism estimator g_n and the adjustment set selected for the treatment mechanism estimator g_n (i..e, the one in the final clever covariate), the empirical variance of the influence curve at the C-TMPLE and the treatment mechanism estimator, and the initial estimator Q_n^0.

Secondly, we re-estimate the treatment effect with the C-TMLE , i.e. based on the regular (non-penalized) log-likelihood, using the same initial estimator Q_n^0 as in the C-TMPLE. Again, we report the estimate, the selected treatment mechanism estimator, its adjustment set, and the empirical variance of the influence curve at the C-TMLE and the treatment mechanism estimator.

We now suggest that the adjustment set of the treatment mechanism selected by the C-TMLE represents the set of confounders one would like to adjust for in order to remove all confounding. These confounders are related to both the outcome and the treatment. On the other hand, the confounders in the adjustment set of the treatment mechanism selected by the penalized C-TMPLE represents the set of confounders we were able to adjust for without incurring too large a penalty w.r.t. variance. That is, the penalized C-TMPLE accepted bias by not adjusting for some extreme confounders, in order to reduce variance.

As a consequence a comparison of the two sets of output for the penalized and regular C-TMLE's is of interest. In particular, we can report the ratio of the variances $var(IC)_{unpenalized}/var(IC)_{penalized}$. A value close to 1 indicates that penalized was able to adjust for all relevant confounders, but a value much larger than 1 indicates the opposite. In addition, we can report the terms in the regular unpenalized treatment mechanism that are not present in the penalized treatment mechanism. We view these left-out confounders as the ones that are responsible for ETA/sparse data bias (maybe alone, or perhaps in combination with other covariates already in the model). In order to interpret the penalized C-TMPLE, we would state that the reported treatment effect was not adjusted for these left-out confounders (which did appear in the regular C-TMLE). Finally, we can also report the increase in efficiency and change in effect estimate when we remove these confounders

(say one by one) from the treatment mechanism estimator of the regular C-TMLE, allowing us to also provide the user a sense of how strong/problematic these left-out confounders are (or how much of a difference it makes to include them versus not).

If the user supplied adjustment set used in the definition of the treatment effect is such that the treatment mechanism estimators in the penalized and regular C-TMLE are almost identical or that the variance estimates are identical, we can feel confident that we have estimated the parameter of interest and can base statistical inference on the estimated variance of its influence curve. That is, in this case the analysis is complete and satisfactory.

On the other hand, if there is a discrepancy, we will point to the problematic confounders that are causing the problem (the ones in treatment mechanism of the regular C-TMLE that are not included in the treatment mechanism of the penalized C-TMLE), and can even estimate at least some of the problematic treatment probabilities to illustrate the lack of identifiability.

This can now be used to define realistic treatment rules that are based on these problematic confounders by never assigning very unlikely treatments, and a corresponding treatment effect. That is, the hardly identifiable static treatment effect is approximated by a nicely identifiable effect of treatment rules that approximate the static treatments. Specifically, the treatment mechanism of the regular C-TMLE implies rules such as $d_1(W)$ that assigns treatment 1 if $g(1|W) > \alpha$, and zero otherwise, and $d_0(W)$ that assigns treatment 0, if $g(0|W) > \alpha$ and zero otherwise, and we would then define as target parameter $EY_{d_1} - EY_{d_0}$.

Finally, we estimate the realistic parameter, and by now the difference between the penalized and collaborative TMLE will be small, so that interpretation and inference is understood. Some fine tuning of the choice of realistic parameter might be required till the penalization is not causing much change anymore.

To summarize, we are suggesting the following general template for data analysis to deal with sparse data bias. 1) The user provides a target parameter and we use the penalized C-TMPLE to provide our best estimate and inference. 2) If necessary, based on the comparison of the penalized and non-penalized C-TMLE we will warn the user about severe degrees of sparse data bias w.r.t. the target parameter, affecting the reliability of the interpretation of the parameter and its inference. 3) We will show the user where the data is sparse. 4) We suggest realistic parameters in the neighborhood of the

original target parameter for treatment that will suffer much less from sparse data bias and can still be nicely interpreted. 5) We rerun the penalized and regular TMLE to determine a new realistic target parameter, in which the effect of the penalization is small, and report its estimate and inference. This new parameter is perhaps even more interesting than the original, since it more closely corresponds to what occurs in the real world. Since the selection of the realistic parameter is merely based on comparison of variances (not based on effect sizes and p-values) the reported statistical inference might still be reliable.

11.4 Parallel consideration multiple initial and second stage estimators in C-TMLE .

We already stressed the following in our description of the C-TMLE template in section 2.

We can imagine having K initial estimates of the Q part of the likelihood, $Q_1^0, ...Q_K^0$. For any one of these we outline above how to construct a sequences of increasingly data-adaptive nuisance parameter estimators, $\{\hat{g}_1, \ldots, \hat{g}_m\}$ and corresponding two-stage estimators. We also can imagine different methods for obtaining such a sequence of nuisance parameter estimators having the property that $\hat{g}_{k+1}$ is more non-parametric than $\hat{g}_k$. Thus, for each initial estimator choice $j \in 1, \ldots, K$ we have multiple two-stage estimators of Q_0, resulting in a matrix of candidate two-stage estimators. We can select among this using the cross-validated (possibly penalized) log-likelihood.

11.5 Not cross-validating the initial estimator when fine tuning the second stage in the C-TMLE.

When selecting between candidate two stage estimators using the same initial estimator one could use the cross-validated log-likelihood that treats the initial estimator as given. In this way, the selector for the second stage will be more agressive and thereby pursue more bias reduction w.r.t. the parameter of interest than the selector based on honest cross-validation. The use of this quasi cross-validated log-likelihood might even be appropriate across different initial estimators as well as long as the initial estimators itself were fine tuned based on cross-validation so that the empirical log-likelihood of the different initial estimators reflect true fit of the data.

References

O. Bembom, M. Petersen, SY Rhee, W.J. Fessel, S.E. Sinisi, R.W. Shafer, and M.J. van der Laan. Biomarker discovery using targeted maximum likelihood estimation: Application to the treatment of antiretroviral resistant HIV infection. *U.C. Berkeley Division of Biostatistics Working Paper Series*, 221, 2007.

O. Bembom, J.W. Fessel, R.W. Shafer, and M.J. van der Laan. Data-adaptive selection of the adjustment set in variable importance estimation. 2008. URL `http://www.bepress.com/ucbbiostat/paper231`.

M. A. Hernan, B. Brumback, and J. M. Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 11(5):561–570, 2000.

J.H. Holland and J.S. Reitman. Cognitive systems based on adaptive algorithms. *SIGART Bull.*, 63:49–49, 1977. ISSN 0163-5719. doi: http://doi.acm.org/10.1145/1045343.1045373.

J. M. Robins and A. Rotnitzky. Comment on the Bickel and Kwon article, "Inference for semiparametric models: Some questions and an answer". *Statistica Sinica*, 11(4):920–936, 2001.

J. M. Robins, A. Rotnitzky, and M.J. van der Laan. Comment on "On Profile Likelihood" by S.A. Murphy and A.W. van der Vaart. *Journal of the American Statistical Association – Theory and Methods*, 450:431–435, 2000.

J.M. Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, 2000a.

J.M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:1393–1512, 1986.

J.M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials (Minneapolis, MN, 1997)*, pages 95–133. Springer, New York, 2000b.

J.S. Sekhon. Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software, Forthcoming*, 2008.

S. Sinisi and M.J. van der Laan. The deletion/substitution/addition algorithm in loss function based estimation: Applications in genomics. *Journal of Statistical Methods in Molecular Biology*, 3(1), 2004.

M.J. van der Laan and J.M. Robins. *Unified methods for censored longitudinal data and causality*. Springer, New York, 2003.

M.J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

M.J. van der Laan, S. Dudoit, and S. Keles. Asymptotic optimality of likelihood-based cross-validation. *Statistical Applications in Genetics and Molecular Biology*, 3, 2004.

M.J. van der Laan, S. Dudoit, and A.W. van der Vaart. The cross-validated adaptive epsilon-net estimator. *Statistics and Decisions*, 24(3):373–395, 2006.

M.J. van der Laan, E. Polley, and A. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(25), 2007. ISSN 1.

A.W. van der Vaart, S. Dudoit, and M.J. van der Laan. Oracle inequalities for multi-fold cross-validation. *Statistics and Decisions*, 24(3):351–371, 2006.

Chapter 5

Randomized Controlled Trials

5.1 *Covariate Adjustment in Randomized Trials with Binary Outcomes: Targeted Maximum Likelihood Estimation*

The following article appears as it was originally published on the University of California, Berkeley Division of Biostatistics Working Paper Series website in 2007, http://www.bepress.com/ucbbiostat/paper215/.

It was later published in *Statistics in Medicine* in 2008, http://www3.interscience.wiley.com/journal/121500450/abstract.

Covariate Adjustment in Randomized Trials with Binary Outcomes: Targeted Maximum Likelihood Estimation

Kelly L. Moore and Mark van der Laan

Abstract

Covariate adjustment using linear models for continuous outcomes in randomized trials has been shown to increase efficiency and power over the unadjusted method in estimating the marginal effect of treatment. However, for binary outcomes, investigators generally rely on the unadjusted estimate as the literature indicates that covariate-adjusted estimates based on logistic regression models are less efficient. The crucial step that has been missing when adjusting for covariates is that one must integrate/average the adjusted estimate over those covariates in order to obtain the marginal effect. We apply the method of targeted maximum likelihood estimation (MLE), as presented in van der Laan and Rubin (2006), to obtain estimators for the marginal effect using covariate adjustment for binary outcomes. We show that the covariate adjustment in randomized trials using logistic regression models can be mapped, by averaging over the covariate(s), to obtain a fully robust and efficient estimator of the marginal effect, which equals the targeted maximum likelihood estimator (MLE). We present simulation studies that show the targeted MLE increases efficiency and power over the unadjusted method, particularly for smaller sample sizes, even when the regression model is mis-specified.

1 Introduction

Suppose we observe n independent and identically distributed observations of the random vector $O = (W, A, Y) \sim p_0$, where W is a vector of baseline covariates, A is the treatment of interest and $Y = \{0, 1\}$ is the binary outcome of interest, and p_0 denotes the density of O. Causal effects are based on a hypothetical full data structure $X = ((Y_a : a \in \mathcal{A}), W)$ containing the entire collection of counterfactual or potential outcomes Y_a for a ranging over the set of all possible treatments $\mathcal{A}$. The observed data structure O only contains a single counterfactual outcome $Y = Y(A)$ corresponding to the treatment that the subject received. The observed data $O = (W, A, Y \equiv Y(A))$ is thus a missing data structure on X with missingness variable A. We denote the conditional probability distribution of treatment A by $g_0(a|X) \equiv P(A = a|X)$. The randomization assumption or coarsening at random assumption states that A is conditionally independent of the full data X given W, $g_0(A|X) = g_0(A|W)$. In a randomized trial in which treatment is assigned completely at random, we have $g_0(A|X) = g_0(A)$. For the sake of presentation, we assume the treatment A is binary and that A is completely randomized as in a typical randomized trial, but our methods are presented so that it is clear how our estimators generalize to observational studies or randomized trials in which $g_0(A|W)$ is known. In the binary A case, $g_0(1) = p(A = 1) = \delta_0$ and $g_0(0) = p(A = 0) = 1 - \delta_0$ and n_1 the number of subjects in treatment group 1 and n_0 the number of subjects in treatment group 0, and $n = n_1 + n_0$. The quantity of interest is causal effect of treatment A on Y, which, for example, can be defined as the risk difference $\psi = E(Y_1) - E(Y_0)$, where Y_1 and Y_0 are the counterfactual outcomes under treatments 1 and 0 respectively. This quantity is typically estimated in randomized trials with the unadjusted estimate

$$\hat{\psi}_1 = \hat{\mu}_1 - \hat{\mu}_0$$

where $\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} I(A_i = 1)Y_i$ and $\hat{\mu}_0 = \frac{1}{n}\sum_{i=1}^{n} I(A_i = 0)Y_i$. An adjusted effect is also sometimes obtained,

$$\hat{\psi}_W = \hat{P}(Y = 1|A = 1, W) - \hat{P}(Y = 1|A = 0, W).$$

Adjusting for baseline covariates and the issues involved has been discussed in Pocock et al. (2002). Although it has been recognized, at least for linear models, i.e. continuous outcomes, that adjusting for covariates increases the precision of the estimate of the marginal causal effect of treatment, investigators are still resistant to adjusting in logistic models and often rely on the unadjusted estimate. This generally appears to be due to confusion as to how to select the covariates and and how to adjust for them (Pocock et al., 2002). In addition, there is a concern that if data-adaptive procedures are used to select the model for $P(Y = 1|A, W)$ that investigators will be tempted to select the model that provides the most favorable results. However, we recommend that as long as the procedure is determined a priori then we can avoid this latter issue. Thus, a black box type data-adaptive procedure, e.g. forward selection, can still be applied as long as the algorithm and candidate covariates are specified a priori. Adjusting for covariates with main terms in linear models, referred to as analysis of covariance (ANCOVA) in randomized trial literature, for the purpose of estimation of the marginal causal effect has been limited to no interaction terms with treatment. When there is such an interaction term, it is often not clear in the literature on analysis of randomized trial data how one uses this conditional model to obtain a marginal effect. However, even in the absence of the interaction term, the increase in precision has not been observed for non-linear models such as the logistic model. In fact, it has actually been reported that the estimates are not in fact made more precise for logistic models (Hernández et al., 2004; Robinson and Jewell, 1991). The crucial step that has been missing when the parameter of interest is the *marginal* causal effect of A on Y, is that when adjusting for covariates W, one must integrate/average the adjusted estimate over those W in order to obtain a marginal effect estimate that is comparable to the unadjusted effect estimate $\hat{\psi}_1$. This method of averaging over W has been referred to as the G-computation formula and is often applied in observational studies when the treatment or exposure has not been assigned randomly (Robins, 1986, 1987). We show that with this additional step of averaging over W, even when the outcome is binary, and even if the regression model is misspecified, we obtain a more efficient estimate in the randomized trial setting. Such an approach allows for interactions between A and W in the model for $P(Y = 1|A, W)$ while still obtaining a marginal effect. We note that the conditional effect may be the parameter of interest in some studies, for example the effect of a drug conditional on age, and thus the investigator does not want to average over age. In this paper we focus only on the marginal effect and using the covariates W to

obtain the most efficient (precise) estimate of this marginal causal effect in a nonparametric model. We apply the method of targeted maximum likelihood estimation (MLE), as presented in van der Laan and Rubin (2006), to obtain estimators for the marginal effect using covariate adjustment for binary outcomes. This general targeted MLE methodology applies to any estimation problem. In this article we apply it to the risk difference, relative risk and odds ratio, in the context of a randomized trial. Targeted MLE was purposefully named in that maximum likelihood estimators aim for trade-off between bias and variance for the whole density, while the targeted MLE carries out a bias reduction specifically *tailored* for the parameter of interest. Substitution estimators based on standard MLE are often biased with respect to the parameter of interest and do not always converge at a parametric rate. On the other hand, the targeted MLE maps a density estimator (e.g., MLE) into a targeted maximum likelihood estimator (at parameter of interest) so that the corresponding substitution estimator is double robust and locally efficient. That is, this estimator in the randomized trial setting is always consistent and asymptotically linear even when the initial regression estimator for $P(Y|A,W)$ is mis-specified, and is even nonparametrically efficient if the initial estimator is consistent. The general algorithm, provided in van der Laan and Rubin (2006), is to start with initial density estimator, then create a parametric model with parameter ϵ through this given initial density estimator whose scores at $\epsilon = 0$ include the components of the efficient influence curve of the parameter of interest at the given density estimator. It estimates ϵ with MLE of this parametric model and finally updates the new density estimator as the corresponding fluctuation of the given initial density estimator. The algorithm can be iterated until convergence. However in many examples convergence is achieved in a single step as is the case for the examples in this paper. We apply this approach to the estimation of marginal treatment effects including the risk difference, relative risk and odds ratio. The targeted maximum likelihood estimator is a very practically attractive procedure since it can be achieved by simply adding a covariate to an initial estimate of the regression $P(Y = 1|A,W)$. The corresponding coefficient ϵ for this new covariate can be estimated with standard software and thus has a straightforward implementation. We show that for the logistic regression model for $P(Y = 1|A,W)$, that this covariate is none other than a linear combination of the treatment variable A so that it follows that the targeted MLE coincides with the standard G-computation ML estimator. This is not always true as we show that these two estimators differ when the treatment mechanism is estimated from the data, which results in an additional efficiency gain. We appeal to estimating function methodology (van der Laan and Robins, 2003) and observe that since the targeted MLE solves the efficient influence curve estimating equation it is double robust and (locally) efficient. That is, the targeted MLE is always consistent and asymptotically linear (thus the standardized estimator is asymptotically normally distributed with specified variance), even if the initial estimate for $P(Y = 1|A,W)$ is misspecified. In the case that the initial estimate for $P(Y = 1|A,W)$ is asymptotically consistent the targeted MLE is asymptotically efficient for the nonparametric model. In section 2 we provide a brief overview of methods for covariate adjustment that have been proposed in literature. In section 3 we present the targeted maximum likelihood estimators for three marginal variable importance parameters: the risk difference, relative risk and odds ratio. We show that for each of these three parameters, using a logistic regression model, the targeted MLE is achieved in a single step. We also provide an alternative to the logistic regression model for the relative risk parameter that is the relative risk regression model and provide the corresponding targeted MLE estimator. We also address missing data on the outcome of covariates, and estimation of the treatment mechanism. Section 4 provides testing and inference for the targeted MLE. In section 5 we present simulation studies that demonstrate the performance of the targeted MLE. Finally we conclude with a discussion in section 6.

2 Current Methods for Obtaining Covariate-Adjusted Estimates

Suppose we observe $O = (W, A, Y)$ as above except the outcome Y is now continuous. Let the parameter of interest be the marginal effect of A on Y, $\psi = E(Y_1) - E(Y_0)$. For a continuous outcome Y, $Q(A,W) = E(Y|A = 1, W)$ is typically obtained using a linear regression model such as,

$$\hat{Q}(A,W) = \hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 W.$$

In this setting, $\hat{\beta}_1$ coincides with and has been shown to be at least as precise as the unadjusted estimate $\hat{\psi}_1$. In particular, the increase in precision occurs when the correlation between the covariate(s) and outcome

is strong (Assmann et al., 2000). However, when $Q(A,W)$ is estimated as

$$\hat{Q}(A,W) = \hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 W + \hat{\beta}_3 AW,$$

then $\hat{\beta}_1$ no longer coincides with $\hat{\psi}_1$. In this case, to obtain the *marginal* effect, one must integrate out or average over the covariate(s) W. The G-computation estimator introduced in Robins (1986) and Robins (1987) is an estimator that does indeed average over W and thus give a marginal effect,

$$\hat{\psi}_{Gcomp} = \frac{1}{n}\sum_{i=1}^{n} \hat{Q}(1,W_i) - \hat{Q}(0,W_i).$$

When $\hat{Q}(A,W)$ is estimated with a linear model, and it does not contain any interaction terms, then $\hat{\psi}_{Gcomp} = \hat{\beta}_1$. The G-computation estimator is not limited to a linear model for $Q(A,W)$ when estimating the treatment effect, for example, when the outcome is binary, one could use a logistic regression model to estimate $Q(A,W)$ and use the G-computation formula to obtain the estimated risk difference. However, even in the absence of interaction terms, $\hat{\psi}_{Gcomp}$ is not necessarily equivalent to the estimate obtained from the logistic regression model. Based on estimating function methodology, the Double Robust (DR) estimator has been provided in van der Laan and Robins (2003); Neugebauer and van der Laan (2005); Robins (2000); Robins and Rotnitzky (2001). Consistency of the DR estimator relies on consistent estimation of the treatment mechanism *or* the model for $Q(A,W)$. When the treatment is randomized, as in a randomized trial, the treatment mechanism is always known and thus the DR estimator is always consistent, i.e. even when $Q(A,W)$ is mis-specified.

It was shown in Scharfstein et al. (1999) (p. $1140-1141$) that to obtain a DR estimate of the difference in two mean outcomes, one can extend a parametric model for $Q(A,W)$ by adding the 2-dimensional covariate $\left(\frac{I(A=1)}{g(1|W)}, \frac{I(A=0)}{g(0|W)}\right)$, where in the randomized trial setting, $g(1|W) = \delta$ and $g(0|W) = 1-\delta$. In section 3.4, under the framework of targeted MLE, we show that for this same additive effect the targeted maximum likelihood algorithm that targets both parameters $(P(Y_0=1), P(Y_1=1))$ also adds these two covariates, the first for $P(Y_1=1)$ and one for $P(Y_0=1)$, so that any function of these two parameters is estimated in a targeted manner. This targeted MLE still differs from the proposal in Scharfstein et al. (1999) by fixing the initial regression, which can thus also represent a data adaptive machine learning fit, and simply estimating the coefficients for the additional covariates. The proposed estimator of Scharfstein et al. (1999) does not fix the initial regression but fits all coefficients for the parametric regression and the additional covariates simultaneously. This distinction in fixing the initial regression is important in that it allows one to apply data adaptive algorithms for the initial estimate and simply update the estimate with the targeting step. This is in contrast to the procedure proposed in Bang and Robins (2005) and Scharfstein et al. (1999) which appears to rely on a parametric estimate for the regression. In Bang and Robins (2005), it is stated that when the initial model for $Q(A,W)$ is correct, then one can obtain a more efficient DR estimate by adding the 1-dimensional covariate $\frac{I(A=1)}{g(1|W)} - \frac{I(A=0)}{g(0|W)}$. This covariate is equivalent to the targeted MLE covariate targeting the risk difference effect $P(Y_1=1) - P(Y_0=1)$. This covariate satisfies the condition of the targeting fluctuation that the score of the initial density $\hat{p}^0$ at $\epsilon = 0$ must include the efficient influence curve at $\hat{p}^0$. Again, the targeted maximum likelihood procedure fixes the initial regression and then estimates the coefficient for the additional covariate as opposed to the proposal in Bang and Robins (2005) where all coefficients for the parametric regression and the additional covariate are fit simultaneously. We note that the covariate that is added in the targeted maximum likelihood algorithm is specific to the parameter one is estimating and thus differs when the parameter of interest is the relative risk or odds ratio as shown in sections 3.2 and 3.3. In section 3.1.1 we provide the relation between the DR, targeted MLE and G-computation estimator and the circumstances in which they coincide.

In Tsiatis et al. (2008), the DR estimator is applied to estimate the marginal effect where the authors recommend estimating two regression models separately: $Q_1(1,W) = E(Y|A=1,W)$ is obtained using only the subpopulation of individuals for whom $A=1$ and $Q_2(0,W) = E(Y|A=0,W)$ is obtained using only the subpopulation of individuals for whom $A=0$. This was proposed so that two different analysts could independently select these models to prevent the analysts from selecting the model providing the most favorable results. Another possibility is to select one model $Q(A,W) = E(Y|A,W)$ using the whole sample pooled together. When the procedure for selecting $Q(A,W)$ is specified a priori this additional

step of estimating $Q_1(1,W)$ and $Q_2(0,W)$ is not necessary. The method provided by Tsiatis et al. (2008) is limited to when the parameter of interest of the marginal effect $E(Y_0) - E(Y_1)$. However, when the outcome is binary, investigators are often also interested in not only the risk difference $E(Y_0) - E(Y_1) = P(Y_1 = 1) - P(Y_0 = 1)$, but the relative risk and odds ratios. Covariate adjustment in logistic regression models for binary outcomes has been studied in literature. However it does not appear that any method for covariate adjustment has been proposed to obtain *marginal* estimates for such parameters. Thus, current applications of logistic regression models provide conditional effects. These conditional models have been shown to *reduce* precision in the estimated effect. In Robinson and Jewell (1991), it was observed that adjusting for covariates in logistic regression models leads to an increase in power due to the fact that estimates of the treatment effect in the conditional logistic models are further away from the null even though standard errors were larger for the adjusted effects. Hernández et al. (2004) also demonstrated this fact using using simulation studies and observed that the increase in power was related to the correlation between the covariate and the outcome. The simulations included only a single covariate and no interactions between the covariate and treatment. Assmann et al. (2000) also indicated similar results in logistic regression models in that odds ratios were generally further away from the null but the standard errors were larger than the unadjusted estimates. It appears that in general, when adjusting for covariates in a logistic regression model, the standard error provided by the software, i.e. standard maximum likelihood procedures, is the standard error used by the investigator although it is often not explicitly stated (Belda et al., 2005; Frasure-Smith et al., 1997; van der Horst et al., 1997; Randolph et al., 2002). When adjusting for covariates in randomized trials using logistic regression, often the investigator is interested in a conditional effect identified by continuous covariates in which case this may be an appropriate approach. We focus on the targeted MLE method for covariate adjustment that provides inference for the marginal (unconditional) effect. However, note that this method can be applied to different subgroups defined by categorical or discrete valued covariates by simple stratification.

3 Targeted Maximum Likelihood Estimation of Marginal Variable Importance: Risk Difference, Relative Risk and Odds Ratio

In this section we present the targeted MLE method for adjusting for covariates when the outcome is binary with the following 3 parameters: risk difference, relative risk and odds ratio.

3.1 Risk Difference

We now provide the targeted MLE for the risk difference $P(Y_1 = 1) - P(Y_0 = 1)$. Let $O = (W, A, Y) \sim p_0$ and $\mathcal{M}$ be the class of all densities of O with respect to an appropriate dominating measure: so $\mathcal{M}$ is nonparametric up to possible smoothness conditions. Consider this non-parametric model for p_0 and let

$$P_0 \to \Psi(p_0) = E_{p_0}(P(Y|A=1,W) - P(Y|A=0,W))$$

be the parameter of interest. This parameter is pathwise differentiable at p_0 with efficient influence curve,

$$\begin{aligned} D(p_0) &= \frac{I(A=1)}{\delta_0}(Y - Q_0(1,W)) - \frac{I(A=0)}{(1-\delta_0)}(Y - Q_0(0,W)) + \\ &\quad + Q_0(1,W) - Q_0(0,W) - \Psi(p_0) \end{aligned}$$

where $Q_0(A,W) = P(Y=1|A,W)$ and $\delta_0 = P(A=1)$ (see e.g., van der Laan and Robins (2003)). Since the model is non-parametric, this is also the only influence curve. Following the strategy of van der Laan and Rubin (2006), the efficient influence curve $D(p_0)$ can be decomposed as,

$$\begin{aligned} D(p_0) &= D(p_0) - E(D(p_0)|A,W) + E(D(p_0)|A,W) - E(D(p_0)|W) + \\ &\quad + E(D(p_0)|W) - E(D(p_0)) \end{aligned}$$

Let, $D_1(p_0) = D(p_0) - E(D(p_0)|A,W)$, $D_2(p_0) = E(D(p_0)|A,W) - E(D(p_0)|W)$ and $D_3(p_0) = E(D(p_0)|A,W) - E(D(p_0))$. Then, $D_1(p_0)$ is a score for $p(Y|A,W)$, $D_2(p_0)$ is a score for $g_0(A|W)$ and $D_3(p_0)$ is a score for the marginal probability distribution $p(W)$ of W. Note that in this randomized trial setting, $g_0(A|W) = g_0(A) = \delta_0^A(1-\delta_0)^{(1-A)}$.

Consider an initial density estimator $\hat{p}^0$ of the density p_0 of O identified by a regression fit $\hat{Q}^0(A,W)$, marginal distribution of A identified by $\hat{\delta} = \frac{1}{n}\sum_{i=1}^n A_i$, the marginal distribution of W being the empirical probability distribution of $W_1, ..., W_n$, and A being independent of W. Since Y is binary, we have the following density,

$$\hat{p}^0(Y|A,W) = (\hat{Q}^0(A,W))^Y (1 - \hat{Q}^0(A,W))^{1-Y}$$

where,

$$\hat{Q}^0(A,W) = \frac{1}{1 + \exp -\hat{m}^0(A,W)}$$

for some function $\hat{m}^0$. Now, consider the parametric submodel through $\hat{p}^0$ indexed by parameter ϵ,

$$\hat{p}^0(\epsilon)(Y|A,W) = (\hat{Q}^0(\epsilon)(A,W))^Y (1 - \hat{Q}^0(\epsilon)(A,W))^{1-Y}$$

where $\hat{Q}^0(\epsilon)(A,W)$ is given by the logistic regression model,

$$\hat{Q}^0(\epsilon)(A,W) = \frac{1}{1 + \exp -(\hat{m}^0(A,W) + \epsilon h(A,W))}$$

with an extra covariate $h(A,W)$, which needs to be chosen so that the score of ϵ at $\epsilon = 0$ includes the efficient influence curve component $D_1(p^0)$ (see van der Laan and Rubin (2006)). The required choice h will be specified below. We estimate ϵ with the maximum likelihood estimator $\hat{\epsilon} = \arg\max_\epsilon \sum_{i=1}^n \log \hat{Q}^0(\epsilon)(A_i, W_i)$. The score for this logistic regression model at $\epsilon = 0$ is given by,

$$\frac{d}{d\epsilon_1} \log p^0(\epsilon)(A,W)\Big|_{\epsilon=0} = h(A,W)(Y - \hat{Q}^0(A,W))$$

We now set the score equal to the part of the efficient IC for $p(Y|A,W)$, that is D_1, at $\hat{p}^0$ to obtain,

$$h(A,W)(Y - \hat{Q}^0(A,W)) = (Y - \hat{Q}^0(A,W)) \left(\frac{I(A=1)}{\hat{\delta}} - \frac{I(A=0)}{(1-\hat{\delta})} \right).$$

This equality in $h(A,W)$ is solved by

$$h(A,W) = \frac{I(A=1)}{\hat{\delta}} - \frac{I(A=0)}{(1-\hat{\delta})}.$$

Thus, the covariate that is added to the logistic regression model $\hat{Q}^0(A,W)$ is none other than a linear combination of A and an intercept only. Thus, if $\hat{m}^0(A,W)$ includes the main term A and the intercept, then $\hat{\epsilon} = 0$, and the targeted MLE for $Q_0(A,W)$ is given by $\hat{Q}^0(A,W)$ itself. In other words, the targeted MLE for ψ_0 is given by the standard G-computation estimator

$$\hat{\psi}_{RD-tMLE} = \frac{1}{n}\sum_{i=1}^n \hat{Q}^0(1,W_i) - \hat{Q}^0(0,W_i).$$

3.1.1 Relation between Targeted MLE, DR and G-computation Estimators

The efficient influence curve $D(p_0)$ can be represented as an estimating function in ψ indexed by Q and g, $D(p_0) = D(Q_0, g_0, \Psi(p_0))$. In this randomized trial setting, $g_0 = \delta_0^A(1-\delta)^{1-A}$. The DR estimate is the solution to the corresponding estimating equation in ψ, $\frac{1}{n}\sum_{i=1}^n D(\hat{Q}^0(A_i,W_i), \hat{\delta}, \psi) = 0$ and is given by,

$$\begin{aligned}\hat{\psi}_{DR} &= \frac{1}{n}\sum_{i=1}^n \frac{I(A_i=1)}{\hat{\delta}}(Y_i - \hat{Q}^0(1,W_i)) - \frac{1}{n}\sum_{i=1}^n \frac{I(A_i=0)}{1-\hat{\delta}}(Y_i - \hat{Q}^0(0,W_i)) + \\ &+ \frac{1}{n}\sum_{i=1}^n \hat{Q}^0(1,W_i) - \frac{1}{n}\sum_{i=1}^n \hat{Q}^0(0,W_i),\end{aligned}$$

where $\hat{\delta} = \frac{1}{n}\sum_{i=1}^{n} A_i$. In the logistic regression fit, $\log(\frac{\hat{Q}(A,W)}{1-\hat{Q}(A,W)}) = \hat{\alpha}X$, where $X = (1, A, W)$, the MLE $\hat{\alpha}$ solves the score equations given by,

$$0 = \sum_{i=1}^{n} X_{ij}(Y_i - \hat{Q}(A_i, W_i)),$$

for $j = 1, ..., p$. The linear span of scores includes the covariate,

$$x_j = \frac{I(A=1)}{\hat{\delta}} - \frac{I(A=0)}{1-\hat{\delta}},$$

when A and an intercept are included in X. Thus, it follows that

$$0 = \frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i=1)}{\hat{\delta}}(Y_i - \hat{Q}^0(1, W_i)) - \frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i=0)}{1-\hat{\delta}}(Y_i - \hat{Q}^0(0, W_i)).$$

Hence,

$$\hat{\psi}_{DR} = \frac{1}{n}\sum_{i=1}^{n} \hat{Q}(1, W_i) - \frac{1}{n}\sum_{i=1}^{n} \hat{Q}(0, W_i) = \hat{\psi}_{Gcomp} = \hat{\psi}_{RD-tMLE}$$

Thus in this quite general scenario, we have that the double robust estimator, the G-computation estimator, and the targeted MLE, all reduce to the same estimator.

3.2 Relative Risk

We now consider the parameter

$$P_0 \to \Psi(p_0) = \frac{E_{p_0}(P(Y|A=1,W))}{E_{p_0}(P(Y|A=0,W)))} = \frac{\mu_1}{\mu_0}$$

Note that under the assumptions listed above for the risk difference, this parameter can be interpreted as the causal relative risk, $\psi_0 = \frac{E(Y_1)}{E(Y_0)}$.

We can derive the efficient influence curve of this parameter using the delta method since we know the efficient influence curve for μ_1 and μ_0. Let $a = \mu_0$ and $b = \mu_1$, so $\psi_0 = \frac{b}{a}$. Then, $\frac{d}{db}\left(\frac{b}{a}\right) = \frac{1}{a}$ and $\frac{d}{da}\left(\frac{b}{a}\right) = -\left(\frac{b}{a^2}\right)$. Thus, the efficient influence curve is given by,

$$\begin{aligned} D(p_0) &= \frac{1}{\mu_0}\left(\frac{I(A=1)}{\delta_0}(Y - Q_0(1,W)) + Q_0(1,W) - \mu_1\right) - \\ &\quad -\frac{\mu_1}{\mu_0^2}\left(\frac{I(A=0)}{(1-\delta_0)}(Y - Q_0(0,W)) + Q_0(0,W) - \mu_0\right) \\ &= \frac{1}{\mu_0}\left(\frac{I(A=1)}{\delta_0}(Y - Q_0(1,W)) + Q_0(1,W)\right) - \\ &\quad \frac{\mu_1}{\mu_0^2}\left(\frac{I(A=0)}{(1-\delta_0)}(Y - Q_0(0,W)) + Q_0(0,W)\right) \end{aligned}$$

We consider two models for the targeted MLE of the relative risk: logistic regression model and the relative risk regression model. In order to find the covariate $h(A, W)$ that is added to the regression model, we note the following equality given in van der Laan and Robins (2003),

$$V(Y, A, W) = (V(1, A, W) - V(0, A, W))(Y - Q(A, W)), \quad (1)$$

if V is a function with conditional mean 0 given A and W. We apply this equality to $D(p_0) = V(Y, A, W)$ to obtain $h(A, W)$.

3.2.1 Submodel 1: Logistic Regression Model

Let $\hat{p}^0(\epsilon_1)$ be the logistic regression fit with an extra covariate extension $\epsilon_1 h(A,W)$. Based on (1) we can immediately observe that the covariate $h(A,W)$ added to the logistic regression is $V(1,A,W)-V(0,A,W)$ since,

$$\begin{aligned}\frac{d}{d\epsilon}\log \hat{p}^0(\epsilon)(A,W)\Big|_{\epsilon=0} &= h(A,W)(Y-\hat{Q}^0(A,W)) \\ &= (V(1,A,W)-V(0,A,W))(Y-\hat{Q}^0(A,W))\end{aligned}$$

Thus, evaluating $D(\hat{p}_0)$ at $Y=1$ and $Y=0$ gives,

$$h(A,W)=\frac{1}{\mu_0}\frac{I(A=1)}{\hat{\delta}}-\frac{\mu_1}{\mu_0^2}\frac{I(A=0)}{(1-\hat{\delta})}.$$

Again, as in the risk difference, the covariate that is added to $\hat{Q}^0(A,W)$ is a function of A only and thus $\hat{\epsilon}=0$ and the targeted MLE for $Q_0(A,W)$ is given by $\hat{Q}^0(A,W)$. The targeted MLE for the relative risk is given by,

$$\hat{\psi}_{RR-tMLE}=\frac{\frac{1}{n}\sum_{i=1}^n \hat{Q}^0(1,W_i)}{\frac{1}{n}\sum_{i=1}^n \hat{Q}^0(0,W_i)}.$$

3.2.2 Submodel 2: Relative Risk Regression

As an alternative to using a logistic fit $Q^0(A,W)$ for $Q(A,W)$, we can instead use a relative risk regression fit,

$$\log(\hat{Q}(A,W))=\hat{m}(A,W),$$

and find the corresponding targeted MLE. Consider now the parametric submodel $\hat{p}^0$ indexed by parameter ϵ,

$$\hat{p}^0(\epsilon)(Y|A,W)=(\hat{Q}^0(\epsilon)(A,W))^Y(1-\hat{Q}^0(\epsilon)(A,W))^{1-Y}$$

where $\hat{Q}^0(\epsilon)(A,W)$ is given by the relative risk regression model,

$$\log(\hat{Q}^0)(\epsilon)(A,W)=\hat{m}^0(A,W)+\epsilon h(A,W).$$

The score for this model evaluated at $\epsilon=0$ is given by,

$$\frac{d}{d\epsilon}\log \hat{p}^0(\epsilon)(A,W)\Big|_{\epsilon=0}=\frac{h(A,W)}{1-\hat{Q}^0(A,W)}(Y-\hat{Q}^0(A,W)),$$

and it follows that the covariate added to logistic regression model to obtain the targeted MLE is given by,

$$h(A,W)=\left(\frac{1}{\mu_0}\frac{I(A=1)}{\hat{\delta}}-\frac{\mu_1}{\mu_0^2}\frac{I(A=0)}{(1-\hat{\delta})}\right)(1-\hat{Q}^0(A,W)).$$

Now $\hat{\epsilon}=\arg\max_\epsilon \sum_{i=1}^n \log \hat{Q}^0(\epsilon)(A_i,W_i)$ can be estimated in practice by fitting a relative risk regression in $\hat{m}^0(A,W)$ and $h(A,W)$, fixing the coefficient in front of $\hat{m}^0(A,W)$ to 1 and the intercept to 0. The resulting coefficient for $h(A,W)$ is $\hat{\epsilon}$. In this case, the covariate is no longer simply a function of A and thus $\hat{\epsilon}$ does not necessarily equal 0 and the targeted MLE is no longer achieved in one step but rather iteratively. Now $\hat{Q}^k(A,W)$ is updated as,

$$\log(\hat{Q}^{k+1}(A,W))=\hat{m}^k(A,W)+\hat{\epsilon}h^k(A,W),$$

setting $k=k+1$ and one iterates this updating step.

3.3 Odds Ratio

We now consider the parameter

$$P_0 \to \Psi(p_0) = \frac{E_{p_0}(P(Y|A=1,W))/(1-E_{p_0}(P(Y|A=1,W)))}{E_{p_0}(P(Y|A=0,W))/(1-E_{p_0}(P(Y|A=0,W)))} = \frac{\mu_1/(1-\mu_1)}{\mu_0/(1-\mu_0)}$$

Note that under the assumptions listed above for the risk difference, this parameter can be interpreted as the causal odds ratio, $\frac{E(Y_1)/(1-E(Y_1))}{E(Y_0)/(1-E(Y_0))}$. Again, applying the delta method we can obtain the efficient influence curve for this parameter. Let $a = \mu_0$ and $b = \mu_1$, so $\psi = \frac{b/(1-b)}{a/(1-a)}$. Then, $\frac{d}{db}\left(\frac{b/(1-b)}{a/(1-a)}\right) = \frac{(1-a)}{a(1-b)^2}$ and $\frac{d}{da}\left(\frac{b/(1-b)}{a/(1-a)}\right) = -\left(\frac{b}{a^2(1-b)}\right)$. Thus, the efficient influence curve is given by,

$$\begin{aligned} D(p_0) &= \frac{1-\mu_0}{\mu_0(1-\mu_1)^2}\left(\frac{I(A=1)}{\delta_0}(Y-Q_0(1,W))+Q_0(1,W)-\mu_1\right) - \\ &\quad -\frac{\mu_1}{(\mu_0)^2(1-\mu_1)}\left(\frac{I(A=0)}{(1-\delta_0)}(Y-Q_0(0,W))+Q_0(0,W)-\mu_0\right) \end{aligned}$$

Applying equality (1) to $D(\hat{p}^0)$, we obtain,

$$h(A,W) = \frac{(1-\mu_0)}{\mu_0(1-\mu_1)^2}\frac{I(A=1)}{\hat{\delta}} - \frac{\mu_1}{\mu_0^2(1-\mu_1)}\frac{I(A=0)}{(1-\hat{\delta})}$$

Again, the covariate that is added to the logistic regression model $\hat{Q}^0(A,W)$ is none other than a function of A only and thus $\hat{\epsilon}=0$ and the targeted MLE for $Q_0(A,W)$ is given by $\hat{Q}^0(A,W)$. Thus, the targeted MLE for ψ is given by,

$$\hat{\psi}_{OR-tMLE} = \frac{\left(\frac{1}{n}\sum_{i=1}^n \hat{Q}^0(1,W_i)\right)/\left(1-\frac{1}{n}\sum_{i=1}^n \hat{Q}^0(1,W_i)\right)}{\left(\frac{1}{n}\sum_{i=1}^n \hat{Q}^0(0,W_i)\right)/\left(1-\frac{1}{n}\sum_{i=1}^n \hat{Q}^0(0,W_i)\right)}.$$

3.4 Targeted MLE for the two treatment specific means, and thereby for all parameters.

Consider the odds ratio, as an example. An alternative for targeting the odds ratio is to simultaneously target both μ_1 and μ_0 and simply evaluate the odds ratio from the targeted MLEs of μ_1 and μ_0. This is a straightforward approach where 2 covariate extensions are added to the logistic fit $\hat{Q}^0$,

$$h_1(A,W) = \epsilon_1\frac{I(A=1)}{\hat{\delta}},$$

and,

$$h_2(A,W) = \epsilon_2\frac{I(A=0)}{(1-\hat{\delta})}.$$

Again, if the initial logistic regression fit already includes an intercept and main term A, then $\hat{\epsilon}=0$ so that this targeted MLE $\hat{Q}=\hat{Q}^0(\hat{\epsilon})=\hat{Q}^0$ is not updated. This targeted MLE can now be used to map into a locally efficient estimator of any parameter of μ_0,μ_1 such as the risk difference $\mu_1-\mu_0$, the relative risk μ_1/μ_0 and the odds ratio $\mu_1(1-mu_0)/((1-\mu_1)\mu_0)$.

3.5 Estimating the Treatment Mechanism as well

Even when the treatment mechanism (the way treatment was assigned) is known as it is in a randomized trial, it has been shown that efficiency is increased when estimating it from the data (van der Laan and Robins, 2003). Estimating the treatment mechanism does not add any benefit to the G-computation estimator since it does not use this information. The targeted MLE can however leverage this information

to obtain a more precise estimate of the treatment effect. This can be a particular benefit when the model for $Q(A,W)$ is mis-specified. The targeted MLE is still consistent when $Q(A,W)$ is mis-specified, however, we can gain efficiency when estimating the treatment mechanism in such a case. The treatment mechanism can be estimated from the data using a logistic regression model, for example, $\hat{g}^0(1|W) = \frac{1}{1+\exp(-(\alpha_1 W_1 + \alpha_2 W_2))}$, but one can also augment an initial fit $\hat{g}^0$ with a targeted direction aiming for a maximal gain in efficiency: see van der Laan and Rubin (2006). We present the targeted MLE for the risk difference, however, this can be immediately extended to the relative risk and odds ratio as well. Consider the parametric submodel through $\hat{p_0}$ indexed by parameter ϵ,

$$\hat{p}^0(\epsilon)(Y|A,W) = (\hat{Q}^0(\epsilon)(A,W))^Y (1 - \hat{Q}^0(\epsilon)(A,W))^{1-Y}$$

where $\hat{Q}^0(\epsilon)(A,W)$ is given by the logistic regression model,

$$\hat{Q}^0(\epsilon)(A,W) = \frac{1}{1 + \exp -(\hat{m}^0(A,W) + \epsilon h(A,W))}.$$

Setting the score of this model equal to the part of the efficient influence curve that corresponds with scores for $P(Y|A,W)$, and solving for $h(A,W)$ we obtain the covariate,

$$h(A,W) = \frac{I(A=1)}{\hat{g}^0(1|W)} - \frac{I(A=0)}{\hat{g}^0(0|W)},$$

which is added to the logistic regression $\hat{Q}^0(A,W)$. Again,
$\hat{\epsilon} = \arg\max_\epsilon \sum_{i=1}^n \log \hat{Q}^0(\epsilon)(A_i,W_i)$ can be estimated in practice by fitting a logistic regression in $\hat{m}^0(A,W)$ and $h(A,W)$, fixing the coefficient in front of $\hat{m}^0(A,W)$ to 1 and the intercept to 0. The resulting coefficient $\hat{\epsilon}$ for $h(A,W)$ is no longer necessarily equal to 0. Let the targeted MLE for $Q_0(A,W)$ be given by $\hat{Q}^*(A,W) = \hat{Q}^0(\hat{\epsilon})(A,W)$. The targeted MLE for ψ_0 is then,

$$\begin{aligned}\hat{\psi}_{RD-tMLE2} &= \frac{1}{n}\sum_{i=1}^n \frac{I(A_i=1)}{\hat{g}^0(1|W)}(Y_i - \hat{Q}^*(1,W_i)) - \\ &\quad -\frac{1}{n}\sum_{i=1}^n \frac{I(A_i=0)}{\hat{g}^0(0|W)}(Y_i - \hat{Q}^*(0,W_i)) + \\ &\quad +\frac{1}{n}\sum_{i=1}^n \hat{Q}^*(1,W_i) - \frac{1}{n}\sum_{i=1}^n \hat{Q}^*(0,W_i).\end{aligned}$$

Note that $\hat{Q}^0(A,W)$ is now updated, contrary to the case when we were not estimating the treatment mechanism as in previous subsections.

3.6 Missing Data

Here we provide the targeted MLE for the case that the outcome Y is subject to missingness that can be informed by the baseline covariates W. In such a case the missingness cannot be ignored as it can lead to biased estimates as treatment groups are no longer balanced with respect to the covariates. Let C represent the indicator whether or not the outcome was observed. The observed data can be represented as $O = (W,A,C,CY) \sim p_0$ and the full data is given by $X = ((Y_a : a \in \mathcal{A}), W)$. We assume that the conditional distribution of the joint censoring variable (A,C) given X satisfies coarsening at random (CAR), i.e. $g_0(A,C|X) = g_0(A,C|W)$. Let

$$P_0 \to \Psi(p_0) = E_{p_0}(P(Y|A=1,W) - P(Y|A=0,W))$$

be the parameter of interest. We wish to estimate the risk difference with the targeted MLE. The efficient influence curve is given by,

$$\begin{aligned} D(p_0) &= \frac{I(A=1)}{g_0(1,1|W)}(Y - Q_0(1,1,W)) - \\ &\quad - \frac{I(A=0)}{(g_0(0,1|W))}(Y - Q_0(0,1,W)) + \\ &\quad + Q_0(1,1,W) - Q_0(0,1,W) - \Psi(p_0), \end{aligned}$$

where $g_0(A=1,c|W) = \delta_0 g(c|A=1,W)$ and $g_0(A=0,c|W) = (1-\delta_0)g(c|A=0,W)$. We now present the analogue to the derivation of the targeted MLE for ψ_0. Consider the parametric submodel through $\hat{p}^0$ indexed by parameter ϵ,

$$\hat{p}^0(\epsilon)(Y|A,C=1,W) = (\hat{Q}^0(\epsilon)(A,C=1,W))^Y (1 - \hat{Q}^0(\epsilon)(A,C=1,W))^{1-Y}$$

where $\hat{Q}^0(\epsilon)(A,C=1,W)$ is given by the logistic regression model,

$$\hat{Q}^0(\epsilon)(A,C=1,W) = \frac{1}{1+\exp -(\hat{m}^0(A,C=1,W) + \epsilon h(A,C=1,W))}.$$

At $C=0$, the likelihood of $P(Y \mid A,C,W)$ provides as contribution a factor 1, which can thus be ignored. The score for this logistic regression model at $\epsilon = 0$ is given by,

$$\frac{d}{d\epsilon} \log p^0(\epsilon)(A,C,W)\Big|_{\epsilon=0} = I(C=1)h(A,C=1,W)(Y - \hat{Q}^0(A,C=1,W))$$

We now set this score equal to the component of the efficient influence curve which equals a score for $P(Y|A,C=1,W)$, at $\hat{p}^0$, to obtain the equality

$$\begin{aligned} &h(A,C=1,W)(Y - \hat{Q}^0(A,C=1,W)) \\ &\quad = (Y - \hat{Q}^0(A,C=1,W)) \left(\frac{I(A=1)}{\hat{g}(1,1|W)} - \frac{I(A=0)}{\hat{g}(0,1|W)} \right). \end{aligned}$$

Solving for $h(A,C=1,W)$ we obtain,

$$h(A,C=1,W) = \frac{I(A=1)}{\hat{g}(1,1|W)} - \frac{I(A=0)}{\hat{g}(0,1|W)}.$$

The estimate of ϵ given by $\hat{\epsilon} = \arg\max_\epsilon \sum_{i=1}^n I(C_i=1) \log \hat{Q}^0(\epsilon)(A_i,W_i)$. Now the logistic regression fit $\hat{Q}^0(Y|A,C=1,W)$ can be updated by adding as covariate $h(A,C=1,W)$ to obtain the targeted MLE $\hat{Q}^*(Y|A,C=1,W)$ for $Q_0(A,C=1,W)$ based on all observations with $C_i=1$. The estimate for $P(C=1|A=0,W)$ as required to calculate the extra covariate $h(A,W)$ can be obtained by using a logistic regression model selected either data-adaptively or using a fixed pre-specified model for C conditional on $W, A=0$. The targeted MLE for ψ_0 is given by,

$$\hat{\psi}_{RD-tMLE} = \frac{1}{n} \sum_{i=1}^n \hat{Q}^*(1,1,W_i) - \hat{Q}^*(0,1,W_i).$$

We note that the targeted MLE for missing covariate values is derived in exactly the same manner.

4 Testing and Inference

Let $\hat{p}^*$ represent the targeted MLE of p_0. One can construct a Wald-type 0.95-confidence interval based on the estimate of the efficient influence curve, $\hat{IC}(O) = D(\hat{p}^*)$. That is, one can estimate the asymptotic variance of $\sqrt{n}(\hat{\psi} - \psi_0)$ with

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \hat{IC}^2(O_i).$$

The corresponding asymptotically conservative Wald-type 0.95-confidence interval is defined as $\psi_n \pm 1.96\frac{\hat{\sigma}}{\sqrt{n}}$. The null hypothesis $H_0 : \psi_0 = 0$ can be tested with the test statistic

$$T_n = \frac{\psi_n}{\frac{\hat{\sigma}}{\sqrt{n}}},$$

whose asymptotic distribution is $N(0,1)$ under the null hypothesis. We note that this estimate of the asymptotic variance is conservative even if $\hat{Q}^0(A,W)$ is inconsistent, and it is actually asymptotically accurate if $\hat{Q}^0(A,W)$ is consistent (see van der Laan and Robins (2003); van der Laan and Rubin (2006)). An alternative recommended approach to obtain a non-conservative estimate of the variance is the bootstrap procedure which will provide asymptotically valid confidence intervals.

5 Simulation Studies

5.1 Simulation 1

In this simulation, the treatment A and outcome Y are binary and W is a 2-dimensional covariate, $W = (W_1, W_2)$. The simulated data were generated according to the following laws:

1. $W_1 \sim N(2,2)$
2. $W_2 \sim U(3,8)$
3. $P(A=1) = \delta_0 = 0.5$
4. $Q_0(A,W) = P(Y=1|A,W) = \frac{1}{(1+\exp(-(kA-5W_1^2+2W_2)))}$

We simulated the data for 2 scenarios based on the value for k in $P(Y=1|A,W)$. In the first scenario, $k=1.2$ and there is a small treatment effect and in the second $k=20$, and there is a larger treatment effect. The risk difference, relative risk and odds ratio were estimated. The true values were given by $P(Y_1=1) = 0.372$, $P(Y_0=1) = 0.352$ and $(RD, RR, OR) = (0.019, 1.055, 1.087)$ for $k=1.2$, $P(Y_1=1) = 0.583$, $P(Y_0=1) = 0.352$ and $(RD, RR, OR) = (0.231, 1.654, 2.570)$ for $k=20$. The parameters were estimated using 4 methods. The first method "Unadjusted" is the unadjusted method of regressing Y on A using a logistic regression model. The second method "Correct" is the targeted maximum likelihood method which is equivalent to the standard G-computation (maximum likelihood) estimator with $\hat{Q}(A,W) = 1/(1+\exp(-(\hat{\alpha}_0 + \hat{\alpha}_1 A + \hat{\alpha}_2 W_1^2 + \hat{\alpha}_3 W_2)))$. The third method "Mis-spec" used a mis-specified fit given by $\hat{Q}(A,W) = 1/(1+\exp(-(\hat{\alpha}_0 + \hat{\alpha}_1 A + \hat{\alpha}_2 W_1)))$. For the fourth method ,"DSA", the estimate $\hat{Q}(A,W)$ was obtained using Deletion/Substitution/Addition (DSA). The DSA algorithm is a data-adaptive model selection procedure based on cross-validation that relies on deletion, substitution, and addition moves to search through a large space of possible functional forms, and is publicly available at http://www.stat.berkeley.edu/ laan/Software/ (Sinisi and van der Laan, 2004). The variable A was forced into the model and the DSA then selects from the remaining covariates. The maximum power set in the DSA algorithm for any term in the model was set to 2, meaning square terms and 2-way interactions were allowed. Standard errors for the targeted MLE were estimated using the estimated influence curve. For the odds ratio simulations, the estimator obtained by extracting the coefficient for A and the corresponding standard error from the logistic regression model fit is labeled "Adjusted". The simulation was run 1000 times for each sample size: $n = 50, 100, 250, 500, 1000$.

For $k=1.2$, W strongly predicts Y and thus the targeted MLE, which adjusts for W results in a large increase in efficiency over the unadjusted method as observed by the relative efficiencies (RE) provided in Table 1. The largest gain in efficiency occurs as expected when $\hat{Q}(A,W)$ is correctly specified followed closely by the DSA method, which in general gives a slightly lower bias and slightly higher variability than the correctly specified model due to overfitting of $\hat{Q}(A,W)$. In the scenario where $k=20$, A is more strongly predictive of Y as compared to W and thus the increase in efficiency is not as marked as when $k=1.2$. The largest increase in efficiency for both values of k occurs for the estimates of the odds ratio. When $\hat{Q}(A,W)$ is mis-specified, there is still a noticeable increase in efficiency showing that it is advised to always adjust for covariates. This is a result of the double robustness of the estimator as discussed in

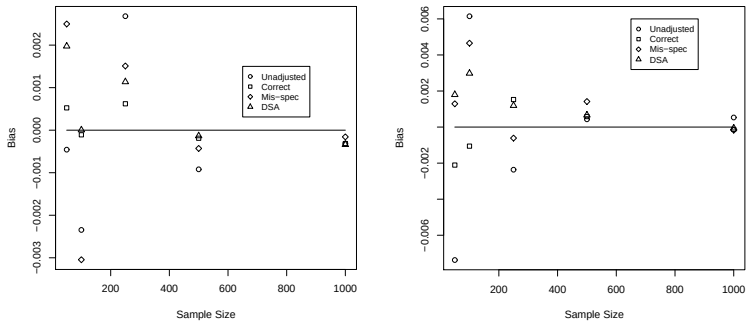
section 2. A significant result is the increase in power of the targeted MLE as evidenced by the proportion of rejected tests. In particular when $k = 1.2$, that is when the effect of A is weaker and more difficult to detect, the increase in power is quite significant. When the sample size is greater than 100, and $k = 20$ the unadjusted performs similar to the targeted MLE estimators with respect to power. Another notable result is that the targeted MLE circumvents the issue of singularity, i.e. Y is perfectly predicted by A and W, that occurs when using the adjusted estimate. In this situation the adjusted estimate is drastically inflated and for this reason, the adjusted results were not included in the bias plots. However, this is not an issue for the targeted MLE. The efficiency gain of the targeted MLE increases as the covariate becomes more predictive. This becomes even more drastic when the covariate is perfectly predictive, whereas the adjusted estimate completely breaks down. For example, in a single run of the simulation for the odds ratio with $k = 1.2$, with $n = 50$, the "Adjusted" model fit gave a coefficient of 25.4 and thus an estimate odds ratio of approximately 10^{11}. The corresponding targeted MLE using this same model gives an estimate of 1.083, noting that the true value is 1.087. This is of particular importance for small sample sizes but still occurs even for large sample sizes as shown in the RE estimates for the "Adjusted" estimate in Table 2. We also note that the bias is almost always positive for the relative risk and odds ratios whereas positive and negative bias occurs for the risk difference.

Table 1: Simulation 1: k=1.2: MSE is Mean Squared Error for Unadjusted Estimate, RE is Relative Efficiency of remaining estimators to Unadjusted MSE and Rej is Proportion of Rejected Tests

	n=50	n=100	n=250	n=500	n=1000
Risk Difference					
Unadjusted MSE	1.8e-02	9.6e-03	3.5e-03	1.9e-03	8.3e-04
Correct RE	5.41	5.01	10.79	12.25	10.95
Mis-spec RE	2.01	2.31	1.95	2.16	2.10
DSA RE	3.38	7.07	10.72	11.99	10.94
Unadjusted Rej	0.06	0.06	0.06	0.08	0.08
Correct Rej	0.18	0.22	0.27	0.41	0.63
Mis-spec Rej	0.09	0.06	0.09	0.11	0.14
DSA Rej	0.09	0.12	0.27	0.42	0.64
Relative Risk					
Unadj MSE	3.0e-01	1.0e-01	3.6e-02	1.5e-02	7.9e-03
Correct RE	9.08	4.07	12.76	12.55	12.34
Mis-spec RE	2.26	2.36	2.10	2.18	2.06
DSA RE	4.09	7.11	12.03	12.22	12.31
Unadjusted Rej	0.04	0.04	0.06	0.06	0.08
Correct Rej	0.10	0.15	0.22	0.37	0.65
Mis-spec Rej	0.05	0.04	0.06	0.07	0.14
DSA Rej	0.03	0.09	0.22	0.37	0.65
Odds Ratio					
Unadj MSE	1.5e+00	3.1e-01	9.5e-02	4.1e-02	2.0e-02
Adjusted RE	9.4e-178	4.8e-251	5.2e-01	5.3e-01	4.2e-01
Correct RE	1.92	0.00	13.49	13.13	12.78
Mis-spec RE	2.97	2.42	2.39	2.28	1.96
DSA RE	7.07	7.05	13.3	12.72	12.57
Unadjusted Rej	0.04	0.06	0.06	0.06	0.09
Adjusted Rej	0.02	0.04	0.04	0.05	0.13
Correct Rej	0.11	0.14	0.21	0.36	0.67
Mis-spec Rej	0.06	0.04	0.04	0.05	0.14
DSA Rej	0.04	0.07	0.21	0.38	0.68

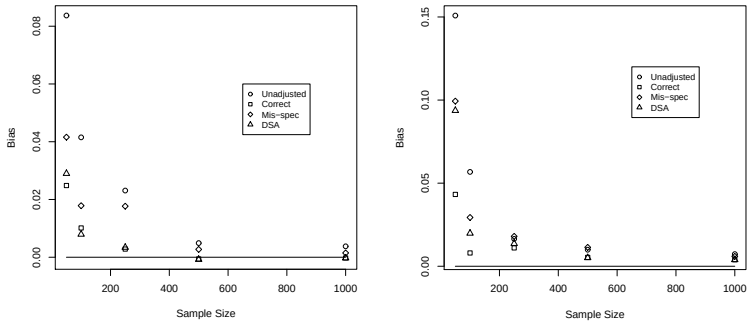
Table 2: Simulation 1: k=20

	n=50	n=100	n=250	n=500	n=1000
Risk Difference					
Unadjusted MSE	2.0e-02	9.2e-03	3.9e-03	1.8e-03	9.9e-04
Correct RE	3.80	3.36	4.16	4.22	4.52
Mis-spec RE	2.25	2.45	2.59	2.49	2.50
DSA RE	2.89	3.86	4.33	4.23	4.52
Unadjusted Rej	0.38	0.68	0.95	1.00	1.00
Correct Rej	0.99	1.00	1.00	1.00	1.00
Mis-spec Rej	0.81	0.97	1.00	1.00	1.00
DSA Rej	0.92	1.00	1.00	1.00	1.00
Relative Risk					
Unadj MSE	5.8e-01	2.0e-01	5.5e-02	2.7e-02	1.4e-02
Correct RE	4.76	4.24	3.63	3.98	4.10
Mis-spec RE	2.01	2.22	2.11	2.11	2.19
DSA RE	2.36	3.34	3.34	3.97	4.09
Unadjusted Rej	0.30	0.61	0.94	1.00	1.00
Correct Rej	0.96	1.00	1.00	1.00	1.00
Mis-spec Rej	0.47	0.92	1.00	1.00	1.00
DSA Rej	0.65	0.98	1.00	1.00	1.00
Odds Ratio					
Unadj MSE	6.9e+00	1.9e+00	6.0e-01	2.4e-01	1.2e-01
Adjusted RE	0.00	0.00	1.7e-17	5.4e-03	4.3e-03
Correct RE	0.00	4.58	2.97	4.87	5.01
Mis-spec RE	2.81	2.79	2.63	2.38	2.58
DSA RE	4.59	4.62	5.27	4.82	5.00
Unadjusted Rej	0.33	0.65	0.96	1.00	1.00
Adjusted Rej	0.44	0.89	1.00	1.00	1.00
Correct Rej	0.94	1.00	1.00	1.00	1.00
Mis-spec Rej	0.25	0.84	1.00	1.00	1.00
DSA Rej	0.52	0.98	1.00	1.00	1.00



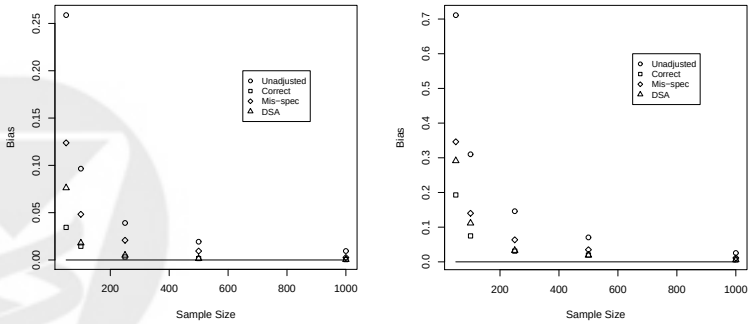
(a) Risk Difference, k=1.2

(b) Risk Difference, k=20



(c) Relative Risk, k=1.2

(d) Relative Risk, k=20



(e) Odds Ratio, k=1.2

(f) Odds Ratio, k=20

Figure 1: Simulation 1: Bias

5.2 Simulation 2: Odds Ratio with Interaction Term

In this simulation, the treatment A and outcome Y are binary and W is a 2-dimensional covariate, $W = (W_1, W_2)$. Here the true causal odds ratio is 0.83. The simulated data were generated according to the following laws:

1. $W_1 \sim N(2, 2)$
2. $W_2 \sim U(3, 8)$
3. $P(A = 1) = \delta_0 = 0.5$
4. $Q_0(A, W) = P(Y = 1|A, W) = \frac{1}{(1+\exp(-(1.2A-5W_1^2+2W_2-5AW_1)))}$

The true values were given by $P(Y_1 = 1) = 0.312$, $P(Y_0 = 1) = 0.352$ and $OR = 0.833$. The same methods used in simulation 1 were used here to estimate the odds ratio. The simulation was run 1000 times for each sample size: $n = 50, 100, 250, 500, 1000$. For the "Mis-spec" targeted MLE, the mis-specified fit was given by $\hat{Q}(A, W) = 1/(1 + \exp(-(\hat{\alpha}_0 + \hat{\alpha}_1 A + \hat{\alpha}_2 W_1)))$. Figure 2 provides a plot of the bias for each of the estimators. The results are similar to odds ratio for simulation 1 in that the bias is positive for all estimators, and thus the odds ratio is over-estimated. Again, even when $\hat{Q}(A, W)$ is mis-specified the bias and MSE are reduced as compared to the unadjusted estimate (Table 3). The DSA, which allows for interactions, shows a significant improvement in terms of bias and MSE. A notable increase in power is again observed for the targeted MLE over the unadjusted method.

Table 3: Odds Ratio, with Interaction

	50	100	250	500	1000
Unadjusted MSE	5.6e-01	1.6e-01	5.9e-02	2.6e-02	1.2e-02
Adjusted RE	0.00	0.00	0.65	0.56	0.38
Correct RE	7.37	1.67	2.22	7.56	7.71
Mis-spec RE	2.78	2.52	2.44	2.60	2.69
DSA RE	5.30	5.69	6.65	7.26	7.68
Unadjusted Rej	0.05	0.07	0.10	0.17	0.31
Adjusted Rej	0.02	0.05	0.13	0.25	0.50
Correct Rej	0.99	0.99	1.00	1.00	1.00
Mis-spec Rej	0.96	1.00	1.00	1.00	1.00
DSA Rej	0.98	1.00	1.00	1.00	1.00

5.3 Simulation 3: Estimating the Treatment Mechanism as well

In this simulation, the treatment mechanism, $\hat{P}(A|W)$ is estimated from the data using a logistic regression model with covariates that are predictive of the outcome Y. The simulated data were generated according to the following laws:

1. $W_1 \sim N(1, 2)$
2. $W_2 \sim U(1, 4)$
3. $W_3 \sim U(0, 20)$
4. $P(A = 1) = \delta_0 = 0.5$
5. $Q_0(A, W) = P(Y = 1|A, W) = \frac{1}{(1+\exp(-(3A-2W_1^2-\log(W_2)+0.5W_3)))}$

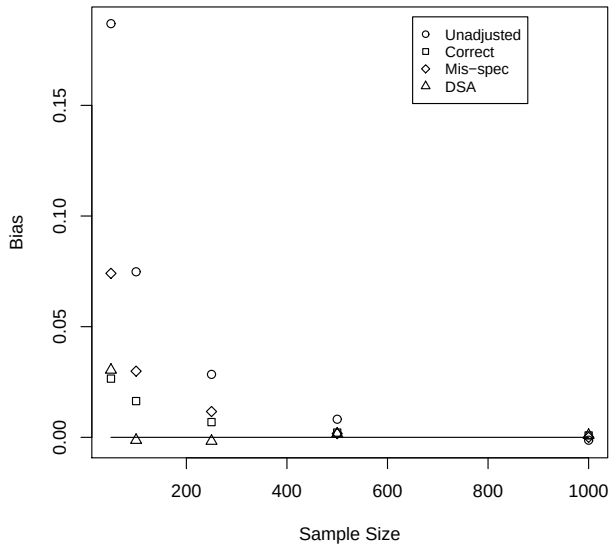


Figure 2: Odds Ratio, Interaction

The true values were given by $P(Y_1 = 1) = 0.569$, $P(Y_0 = 1) = 0.419$ and $RD = 0.150$. The treatment mechanism was estimated with the logistic regression model given by $g(A|W) = 1/(1+\exp(-(\gamma_0+\gamma_1 W_1+\gamma_2 W_2+\gamma_3 W_3)))$. The targeted MLE estimator, represented as "Est tx" in Table 5 and Figure 4, with the estimated treatment mechanism is no longer equivalent to the G-computation estimator. The mis-specified fit for $Q(A, W) = 1/(1+\exp(-(\alpha_0+\alpha_1 A+\alpha_2 W_1)))$ is used as the initial fit and the covariate $h(A, W)$ provided in section 3.4 is then added to this logistic regression. The targeted MLE is then estimated as usual. Thus, we are interested in comparing the mis-specified targeted MLE to the estimated treatment mechanism targeted MLE. Figure 4 shows the bias is reduced and the efficiency is slightly increased when estimating the treatment mechanism. The power was approximately equal for the mis-specified and estimated treatment mechanism targeted MLE. The DSA targeted MLE method again shows a large improvement in efficiency and power over the unadjusted method.

Table 4: Risk Difference, Estimated Tx Mechanism

	50	100	250	500	1000
Unadjusted MSE	2.1e-02	9.4e-03	3.8e-03	1.9e-03	9.9e-04
Correct RE	3.22	3.51	3.91	3.95	4.08
DSA RE	2.65	3.48	3.89	3.94	4.04
Mis-spec RE	1.19	1.18	1.16	1.21	1.20
Est tx RE	1.26	1.30	1.28	1.34	1.29
Unadjusted Rej	0.22	0.34	0.67	0.92	1.00
Correct Rej	0.73	0.90	1.00	1.00	1.00
DSA Rej	0.59	0.90	1.00	1.00	1.00
Mis-spec Rej	0.26	0.42	0.76	0.96	1.00
Est tx Rej	0.23	0.40	0.75	0.96	1.00

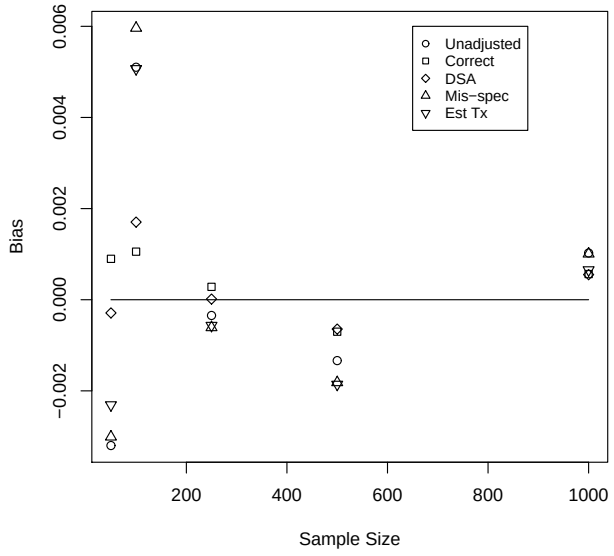


Figure 3: Risk Difference, Estimated Treatment Mechanism

5.4 Efficiency Gain and R^2

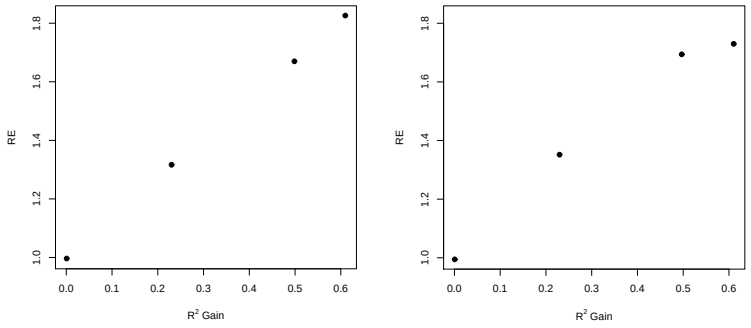
The gain in relative efficiency is related to the gain in the squared multiple correlation coefficient R^2. A covariate predictive of the outcome results in an increase in R^2 in the adjusted model as compared to the unadjusted model. The increase in R^2 results in an increase in efficiency in the targeted MLE. Pocock et al. (2002) discussed the increase in efficiency when adjusting for predictive covariates in linear models. The following simulations show that this also applies to the targeted MLE using logistic regression models. Simulated data were generated according to the following laws:

1. $\sqrt{W} \sim N(2, 2)$
2. $P(A = 1) = \delta_0 = 0.5$
3. $Q_0(A, W) = P(Y = 1|A, W) = \frac{1}{(1+\exp(-(1.2A - cW)))}$

A simulation of sample size $n = 1000$ was run for each $c = \{0, 0.25, 2, 10\}$, that is covariate W is increasingly predictive. The R^2 was estimated in the ordinary least squares sense,

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Q}(A, W))^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}.$$

A gain in R^2 was computed as the difference between R^2 in the covariate adjusted model and the covariate unadjusted model. Figure 5 and 6 depict the relative efficiency to the unadjusted model for the targeted MLE of the odds ratio against the gain in R^2 for the targeted MLE of the odds ratio and risk difference respectively.



(a) Risk Difference (b) Odds Ratio

Figure 4: Efficiency Gain and R^2

5.5 Simulations Discussion

The four simulations were relatively simple scenarios but were useful in demonstrating the following points:

- The targeted MLE shows a clear increase in both efficiency and power over the unadjusted method, even when $Q(A, W)$ is not correctly specified.
- The DSA method for selecting $Q(A, W)$ provides a significant increase in efficiency and power over the mis-specified fixed $Q(A, W)$ method. The average relative efficiencies between these two methods ranged from 1.7 to 3.6 for sample sizes $n = 50$ to $n = 1000$ in our simulations.
- The targeted MLE circumvents the singularity issue that occurs when using the adjusted method of extracting the coefficient from the logistic regression model $Q(A, W)$.

- Interaction terms in the model for $Q(A, W)$ fit entirely into the framework of the targeted MLE.
- Estimating the treatment mechanism provides a further small increase in efficiency over targeting only $Q(A, W)$.

6 Discussion

The targeted MLE provides a general framework that we applied to estimation of the marginal (unadjusted) effect of treatment in randomized trials. We observed that the traditional method of covariate adjustment in randomized trials using logistic regression models can be mapped, by averaging over the covariate(s), to obtain a fully robust and efficient estimator of the marginal effect, which equals the targeted MLE. We demonstrated that the targeted MLE does just this and results in an increase in efficiency and power over the unadjusted method, contrary to what has been reported in the literature for covariate adjustment for logistic regression. The simulation results showed that data-adaptive model selection algorithms such as the DSA, which we used in this paper, or forward selection, when specified a priori should be used. However, we showed that even adjusting by a misspecified regression model results in gain in efficiency and power. Thus, using an a priori specified model, even if it is mis-specified, can increase the power, and thus reduce the sample size requirements for the study. This is particularly important for trials with smaller sample sizes. The targeted MLE framework can also address missing data, either in the outcome as we demonstrated in section 3.5 for the risk difference, but also missingness in covariates and treatment as well for any of the parameters of interest. In these scenarios the targeted MLE covariate may not be as straightforward as those that were presented in this paper, but its derivation is analogue. We focused on logistic and relative risk regression, but the methodology can be extended to any other regression models for $Q(A, W)$. The targeted MLE framework can also be applied to other parameters of interest in randomized trials such as an adjusted effect, for example by age or biomarker, and can also handle survival times as outcomes (see van der Laan and Rubin (2006)).

References

Assmann, S., Pocock, S., Enos, L., and Kasten, L. (2000). Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*, 355(9209):1064–1069.

Bang, H. and Robins, J. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Belda, F., Aguilera, L., Garcia de la Asuncion, J., Alberti, J., Vicente, R., Ferrandiz, L., Rodriguez, R., Company, R., Sessler, D., Aguilar, G., Botello, S., Orti, R., and for the Spanish Reduccion de la Tasa de Infeccion Quirurgica Group (2005). Supplemental Perioperative Oxygen and the Risk of Surgical Wound Infection: A Randomized Controlled Trial. *JAMA*, 294(16):2035–2042.

Frasure-Smith, N., Lespérance, F., Prince, R., Verrier, P., Garber, R., Juneau, M., Wolfson, C., and Bourassa, M. (1997). Randomised trial of home-based psychological nursing intervention for patients recovering from myocardial infarction. *Lancet*, 350:473–479.

Hernández, A., Steyerberg, E., and Habbema, J. (2004). Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *J Clin Epidemiol*, 57(5):454–460.

Neugebauer, R. and van der Laan, M. (2005). Why prefer double robust estimators in causal inference? *Journal of the American Statistical Association*, 129:405–426.

Pocock, S., Assmann, S., Enos, L., and Kasten, L. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine*, 21:2917–2930.

Randolph, A., Wypij, D., Venkataraman, S., Hanson, J., Gedeit, R., Meert, K., Luckett, P., Forbes, P., Lilley, M., Thompson, J., Cheifetz, I., Hibberd, P., Wetzel, R., Cox, P., Arnold, J., for the Pediatric Acute Lung Injury, and Network, S. I. (2002). Effect of Mechanical Ventilator Weaning Protocols on Respiratory Outcomes in Infants and Children: A Randomized Controlled Trial. *JAMA*, 288(20):2561–2568.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math. Modelling*, 7(9-12):1393–1512. Mathematical models in medicine: diseases and epidemics, Part 2.

Robins, J. (1987). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Disease*, 40:139S–161S.

Robins, J. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*. American Statistical Association, Alexandria, VA.

Robins, J. and Rotnitzky, A. (2001). Comment on the Bickel and Kwon article, "Inference for semiparametric models: Some questions and an answer". *Statistica Sinica*, 11(4):920–936.

Robinson, L. and Jewell, N. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review*, 59:227–240.

Scharfstein, D., Rotnitzky, A., and Robins, J. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120.

Sinisi, S. and van der Laan, M. (2004). The deletion/substitution/addition algorithm in loss function based estimation: Applications in genomics. *Statistical Applications in Genetics and Molecular Biology*, 3(1).

Tsiatis, A., Davidian, M., Zhang, M., and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine*, 27(23):4658–4677.

van der Horst, C., Saag, M., Cloud, G., Hamill, R., Graybill, J., Sobel, J., Johnson, P., Tuazon, C., Kerkering, T., Moskovitz, B., Powderly, W., and Dismukes, W. (1997). Treatment of cryptococcal meningitis associated with the acquired immunodeficiency syndrome. national institute of allergy and infectious diseases mycoses study group and aids clinical trials group. *The New England Journal of Medicine*, 337(1):15–21.

van der Laan, M. and Robins, J. (2003). *Unified methods for censored longitudinal data and causality*. Springer, New York.

van der Laan, M. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):Article 11.

5.2 *Selecting Optimal Treatments Based on Predictive Factors*

The following article appears as it was originally published on the University of California, Berkeley Division of Biostatistics Working Paper Series website in 2009, http://www.bepress.com/ucbbiostat/paper244/.

It was later published in the book **Design, Summarization, Analysis & Interpretation of Clinical Trials with Time-to-Event Endpoints**, edited by Karl E. Peace for Chapman and Hall as *Predicting Optimal Treatment Assignment Based on Prognostic Factors in Cancer Patients* in 2009.

Selecting Optimal Treatments Based on Predictive Factors

Eric C. Polley and Mark J. van der Laan

1 Introduction

With the increasing interest in individualized medicine there is a greater need for robust statistical methods for prediction of optimal treatment based on the patient's characteristics. When evaluating two treatments, one treatment may not be uniformly superior to the other treatment for all patients. A patient characteristic may interact with one of the treatments and change the effect of the treatment on the response. Clinical trials are also collecting more information on the patient. This additional information on the patients combined with the state-of-the-art in model selection allows researchers to build better optimal treatment algorithms.

In this chapter we introduce a methodology for predicting optimal treatment. The methodology is demonstrated first on a simulation and then on a phase III clinical trial in neuro-oncology.

2 Predicting Optimal Treatment Based on Baseline Factors

Start with a randomized controlled trial where patients are assigned to one of two treatment arms, $A \in \{0, 1\}$, with $\Pr(A = 1) = \Pi_A$. The main outcome for the trial is defined at a given time point t as $Y = \mathrm{I}(T > t)$ where T is the survival time. For example, the main outcome may be the six-month progression-free rate and T is the progression time. Also collected at the beginning of the trial is a set of baseline covariates W. The baseline covariates may be any combination of continuous and categorical variables.

The baseline covariates can be split into prognostic and predictive factors. Prognostic factors are patient characteristics which are associated with the outcome independent of the treatment given, while predictive factors are patient characteristics which interact with the treatment in their association with the outcome. To determine the optimal treatment, a model for how the predictive factors and treatment are related to the outcome needs to be estimated.

The observed data is $O_i = (W_i, A_i, Y_i = \mathrm{I}(T_i > t)) \sim P$ for $i = 1, \ldots, n$. For now assume Y is observed for all patients in the trial but this assumption is relaxed in the next section. The optimal treatment given a set of baseline variables is found using the W-specific variable importance parameter:

$$\Psi(W) = \mathrm{E}(Y|A = 1, W) - \mathrm{E}(Y|A = 0, W) \tag{1}$$

$\Psi(W)$ is the additive risk difference of treatment A for a specific level of the prognostic variables W. The conditional distribution of Y given W is defined as $\{Y|W\} \sim \text{Bernoulli}(\pi_Y)$. The subscript W is assumed on π_Y and left off for clarity of the notation. Adding the treatment variable A into the conditioning statement we define $\{Y|A = 1, W\} \sim \text{Bernoulli}(\pi_{+1})$ and $\{Y|A = 0, W\} \sim \text{Bernoulli}(\pi_{-1})$. Again the subscript W is dropped for clarity but assumed throughout the paper. The parameter of interest can be expressed as $\Psi(W) = \pi_{+1} - \pi_{-1}$. For a given value of W, $\Psi(W)$ will fall into one of three intervals with each interval leading to a different treatment decision. The three intervals for $\Psi(W)$ are:

1. $\Psi(W) > 0$: indicating a beneficial effect of the intervention $A = 1$.

2. $\Psi(W) = 0$: indicating no effect of the intervention A.

3. $\Psi(W) < 0$: indicating a harmful effect of the intervention $A = 1$.

Knowledge of $\Psi(W)$ directly relates to knowledge of the optimal treatment.

As noted in [1], the parameter of interest can be expressed as:

$$\Psi(W) = \mathrm{E}\left(\left(\frac{\mathrm{I}(A=1)}{\Pi_A} - \frac{\mathrm{I}(A=0)}{1-\Pi_A}\right) Y|W\right). \tag{2}$$

When $\Pi_A = 0.5$, the conditional expectation in equation (2) can be modeled with the regression of $Y(A-(1-A))$ on W. Let $Z = Y(A-(1-A))$ and since A and Y are binary variables:

$$Z = \begin{cases} +1 & \text{if } Y = 1 \ \& \ A = 1 \\ 0 & \text{if } Y = 0 \\ -1 & \text{if } Y = 1 \ \& \ A = 0 \end{cases}$$

The observed values of Z follow a multinomial distribution. The parameter $\Psi(W)$ will be high dimensional in most settings and the components of $\Psi(W)$ are effect modifications between W and the treatment A on the response Y. The parameter can be estimated with a model $\Psi(W) = m(W|\beta)$. The functional form of $m(W|\beta)$ can be specified *a priori*, but since the components of the model represent effect modifications, knowledge of a reasonable model may not be available and we recommend a flexible approach called the super learner (described in the next section) for estimating $\Psi(W)$. In many cases a simple linear model may work well for $m(W|\beta)$, but as the true functional form of $\Psi(W)$ becomes more complex, the super learner gives the researcher

flexibility in modeling the optimal treatment function. With the squared error loss function for a specific model $m(W|\beta)$, the parameter estimates are:

$$\beta_n = \arg\min_{\beta} \sum_{i=1}^{n} (Z_i - m(W_i|\beta))^2 \tag{3}$$

The treatment decision for a new individual with covariates $W = w$ is to treat with $A = 1$ if $m(w|\beta_n) > 0$, otherwise treat with $A = 0$.

A normal super learner model for $m(W|\beta)$ would allow for a flexible relationship between W and Z but these models do not respect the fact that $\Psi(W)$ is bounded between -1 and $+1$. The regression of Z on W does not use the information that the parameter $\Psi(W) = \pi_{+1} - \pi_{-1}$ is bounded between -1 and $+1$. The estimates in equation (3) have a nice interpretation since the model predicts the additive difference in survival probabilities. In proposing an alternative method, we wanted to retain the interpretation of an additive effect measure but incorporate the constrains on the distributions. Starting with the parameter of interest in equation (1) we add a scaling value based on the conditional distribution of Y given W as in:

$$\Psi'(W) = \frac{\mathrm{E}_P(Y|A=1,W) - \mathrm{E}_P(Y|A=0,W)}{\mathrm{E}_P(Y|W)} = \frac{\pi_{+1} - \pi_{-1}}{\pi_Y} \tag{4}$$

Since $\pi_Y = \Pr(Y = 1|W) = \Pr(Z \neq 0|W)$, the new parameter $\Psi'(W) = \mathrm{E}(Z|Z \neq 0, W)$. When we restrict the data to the cases with $Z \neq 0$ (i.e. $Y = 1$) the outcome becomes a binary variable and binary regression methods can be implemented. For example, the logistic regression model:

$$\text{logit}\left(\Pr(Z = 1|Z \neq 0, W)\right) = m'(W_i|\beta) \tag{5}$$

The treatment decision is based on $m'(W_i|\beta_n) > 0$ where β_n is the maximum likelihood estimate for the logistic regression model. With the binary regression setting, we are now incorporating the distribution information in creating the prediction model, but losing information by working on a subset of the data. These trade-offs depend on the probability π_Y and we will evaluate both methods on the trial example below. In the next section we propose a data-adaptive method for estimating $\Psi(W)$.

3 Super Learner

Many methods exist for prediction, but for any given data set it is not known which method will give the best prediction. A good prediction algorithm should be flexible to the true data generating distribution. One such algorithm is the super learner [2]. The super learner is applied to predict the optimal treatment based on the observed data. The super learner algorithm starts with the researcher selecting a set of candidate prediction algorithms (candidate learners). This list of candidate learners should be selected to cover a wide range of basis functions. The candidate learners are selected prior to analyzing the data; selection of the candidates based on performance on the observed data may introduce bias in the final prediction model. A flow diagram for the super learner algorithm is provided in figure 19.1. With the candidate learners selected and the data collected, the initial step is to fit all of the candidate learners on the entire data set and save the predicted values for $\Psi_n(W) = m_j(W|\beta_n)$, where j indexes the candidate learners. The data is then split into V equal sized and mutually exclusive sets as is typically done

for V-fold cross-validation. Patients in the v^{th} fold are referred to as the v^{th} validation set, and all patients not in the v^{th} fold are referred to as the v^{th} training set. For the v^{th} fold, each candidate learner is fit on the patients in the v^{th} training set and the predicted values for $\Psi(W) = m_j(W|\beta_n)$ for the patients in the v^{th} validation set are saved. This process of training the candidate learners on the out of fold samples and saving the predicted values in the fold is repeated for all V folds. The predictions from all V folds are stacked together in a new data matrix X^v. With the prediction data, regress the observed outcome Z on the columns of X^v, which represent the predicted outcomes for each candidate learner. This regression step selects weights for each candidate learner to minimize the cross-validated risk. With the estimates, β_n, from the model $\mathrm{E}(Z|X^v) = m(X|\beta)$ the super learner only saves the weights (β_n) and the functional form of the model. The super learner prediction is then based on combining the predictions from each candidate learner on the entire data set with the weights from the cross-validation step.

4 Extensions for Censored Data

In a prospective trial the data may be subject to right censoring. In both methods above, right censoring leads to the outcome Z being missing. The data structure is extended to include an indicator for observing the outcome. Let C be the censoring time (for individuals with an observed outcome we set $C = \infty$). Define $\Delta = \mathrm{I}(C > t)$. $\Delta = 1$ when the outcome is observed and $\Delta = 0$ when the outcome is missing. The observed data is the set $(W, A, \Delta, Y\Delta)$. For the first method, we propose using the doubly

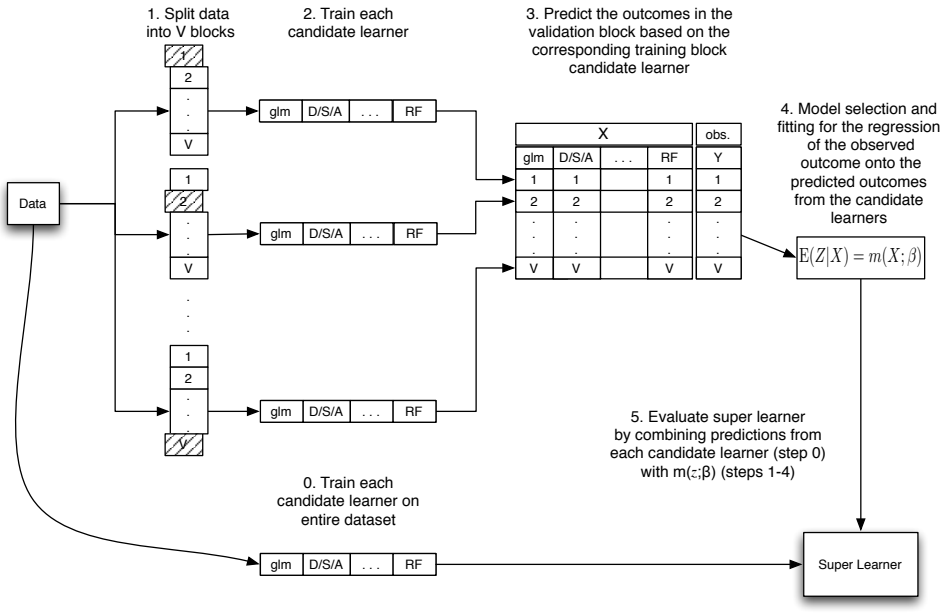


Figure 1: Flow diagram for super learner

robust censoring unbiased transformation [3]. The doubly robust censoring unbiased transformation generates a new variable Z^* which is a function of the observed data but has the additional property:

$$\mathrm{E}\left(Z^*|W, \Delta = 1\right) = \mathrm{E}\left(Z|W\right)$$

The transformation allows estimation of the parameter $\Psi(W)$ by applying the super learner on the uncensored observations with the transformed variable Z^* as the outcome. The doubly robust censoring unbiased transformation is:

$$Z^* = \frac{Z\Delta}{\pi(W)} - \frac{\Delta}{\pi(W)}Q(W) + Q(W), \tag{6}$$

where $\pi(W) = \Pr(\Delta = 1|W)$ and $Q(W) = \mathrm{E}(Z|W, \Delta = 1)$. Both $\pi(W)$ and $Q(W)$ need to be estimated from the data. If either $\pi(W)$ or $Q(W)$ is consistently estimated, then the prediction function $\mathrm{E}(Z^*|W, \Delta = 1) = m(W|\beta_n)$ is an unbiased estimate for the true parameter $\Psi(W)$. The censoring mechanism $\pi(W)$ can be estimated with a logistic regression model or a binary super learner on the entire data set. Similarly, $Q(W)$ may be fit with a linear regression model or a super learner, but on the subset of the data with observed values for Z.

For the second method which relies on modeling $\mathrm{E}(Z|Z \neq 0, W)$, the main feature was the ability to use the knowledge of the distributions to develop a better model. To retain the binary outcome, the doubly robust censoring unbiased transformation will not work. An alternative method for the right censoring which will retain the binary outcome would be inverse probability of censoring weighting. Inverse probability of censoring weights uses the same $\pi(W)$ as above, but does not incorporate the other nuisance parameter $Q(W)$. When applying the binary super learner for $\mathrm{E}(Z|Z \neq 0, W, \Delta = 1)$ the weights $1/\pi(W)$ will be applied for both the candidate learners and the V-fold cross-validation steps. The super learner will minimize the weighted loss function.

5 Simulation Example

We first demonstrate the proposed method on a simulation example where the true value of $\Psi(W)$ is known. The baseline variables were all simulated as normally distributed, $W_j \sim N(0, 1),\ j = 1, \ldots, 10$. The treat-

ment was randomly assigned with $\Pi_A = 0.5$. The true model for the outcome was:

$$\Pr(Y = 1|A, W) = g^{-1}(0.405A - 0.105W_1 + 0.182W_2 + 0.039AW_2 \quad (7)$$
$$+0.006AW_2W_3 - 0.357AW_4 - 0.020AW_5W_6 - 0.051AW_6)$$

Where $g^{-1}(\cdot)$ is the inverse logit function and W_j refers to the j^{th} variable in W. The true model was selected to include interactions between the treatment and some of the baseline variables. With knowledge of the true model for the outcome Y, the true value of $\Psi(W)$ is calculated for every individual.

The first method involves the regression of Z on W. We applied the super learner for $m(W|\beta)$. 10-fold cross validation was used for estimating the candidate learner weights in the super learner. The super learner for the first method included five candidate learners. The first candidate was ridge regression [4]. Ridge regression used an internal cross validation to select the penalty parameter. Internal cross validation means the candidate learner performed a V-fold cross validation procedure within the folds for the super learner. Structurally, when the candidate learner also performs cross validation within the super learner cross validation we have nested cross validation; therefore, we refer to the candidate learner cross validation as internal cross validation. The second candidate was random forests [5]. For the random forest candidate learner, 1000 regression trees were grown. The third candidate was least angle regression [6]. An internal 10-fold cross validation procedure was used to determine the optimal ratio of the L1 norm

Method	R Package	Authors
Adaptive Regression Splines	polspline	Kooperberg
Least Angle Regression	lars	Efron and Hastie
Penalized Logistic	stepPlr	Park and Hastie
Random Forests	randomForest	Liaw and Wiener
Ridge Regression	MASS	Venables and Ripley

Table 1: R Packages for Candidate Learners. R is available at `http://www.r-project.org`

of the coefficient vector compared to the L1 norm of the full least squares coefficient vector. The fourth candidate was adaptive regression splines for a continuous outcome [7]. The final candidate was linear regression. Table 1 contains reference for the R packages implemented for the candidate learners in the super learner.

The prediction model from the super learner is:

$$\Psi_n(W) = -0.01 + 7.24(X_n^{ridge}) + 1.16(X_n^{rf}) - 0.20(X_n^{lars}) - 7.07(X_n^{lm}) - 0.03(X_n^{mars})$$

Where X_n^j is the predicted value for Z based on the j^{th} candidate learner. $j = ridge$ is the ridge regression model. $j = rf$ is the random forests model. $j = lars$ is the least angle regression model. $j = lm$ is the main effects linear regression model. $j = mars$ is the adaptive regression splines model. The largest weights are for ridge regression and the linear regression model. For example, the estimates for the linear regression model is:

$$X_n^{lm} = 0.06 + 0.02W_1 + 0.01W_2 - 0.03W_3 - 0.07W_4 + 0.01W_5 + 0.05W_6 - 0.02W_7 - 0.00W_8 - 0.01W_9 - 0.06W_{10}.$$

The linear regression model has the largest coefficient on W_4, which is the variable with the strongest effect modification with the treatment in the true model (equation (7)). The second largest coefficient is on W_{10} which is a variable unrelated to the outcome. The super learner helps smooth over these errors by having multiple candidate learners. For example, W_{10} has a small coefficient (-0.01) in the ridge regression model. When all the candidates are combined into the final super learner prediction model the spurious effect estimates will often disappear resulting in a better predictor. The third largest coefficient from the linear regression model is on W_6 which is also a strong effect modifier in the true model. To evaluate how the super learner is performing in comparison to the other candidate learners, each candidate learner was also fit as a separate estimate. We looked at two risk values, first the $\mathrm{E}(\Psi_n(W)-Z)^2$ which was minimized by each algorithm. For the simulation, the risk $\hat{E}(Z-\Psi(W))^2 = 0.540$ gives a lower bound for the risk $\mathrm{E}(\Psi_n(W)-Z)^2$. Since the true $\Psi(W)$ is known in the simulation, the risk $\mathrm{E}(\Psi_n(W)-\Psi(W))^2$ was also evaluated. Table 2 contains the risk values for the simulation. The super learner achieved the smallest $\mathrm{E}(\Psi_n(W)-Z)^2$ and is comparable to MARS and LARS on the risk for the true parameter value $\Psi(W)$.

The super learner for the second method included three candidate learners. The first candidate was adaptive regression splines for polychotomous outcomes [8]. The second candidate was the step-wise penalized logistic regression algorithm [9]. The final candidate was main terms logistic regres-

	$\mathrm{E}(\Psi_n(W) - \Psi(W))^2$	$\mathrm{E}(\Psi_n(W) - Z)^2$
Super Learner	0.012	0.544
MARS	0.012	0.549
LARS	0.012	0.549
Ridge	0.026	0.558
Linear Model	0.028	0.559
Random Forests	0.038	0.565

Table 2: Risk for all candidate learners and the super learner

sion. The super learner for the second method is:

$$\Psi'_n(W) = -1.20 + 1.43(X_n^{poly}) - 0.50(X_n^{plr}) + 1.61(X_n^{glm})$$

Where X_n^j is the predicted value for Z based on the j^{th} candidate learner. $j = poly$ is the polyclass adaptive spline model. $j = plr$ is the penalized logistic regression model. $j = glm$ is the main effects logistic regression model.

6 Example of Prediction Model on Clinical Trial

A phase III clinical trial was conducted to evaluate a novel treatment for brain metastasis. The study recruited 554 patients with newly diagnosed brain metastasis and the patients were randomized to receive either standard care ($A = 0$) or the novel treatment ($A = 1$). The researchers were interested in determining an optimal treatment to maximize the probability of surviving 6 months from treatment initiation without progression. Of the 554 patients, 246 are censored prior to 6 months. For the 308 patients with an observed 6 month progression time, 130 progressed or died (42.2%).

In addition to the treatment and event time data, the researchers collected baseline prognostic and predictive factors on every patient. We apply the super learner to estimate a model for selecting the optimal treatment given a patient's baseline factors. A breakdown of the sample size and treatment allocations available for each method is given in table 3.

		A	
	total	0	1
Enrolled	554	275	279
Method 1	308	158	150
Method 2	130	67	63

Table 3: Number of subjects in each treatment arm at enrollment and available for each method.

6.1 Super learner for optimal treatment decisions

Both methods proposed above were applied to the data. The first method looks for a model of Z on W treating Z as a continuous variable. The second method looks for a model of Z on W conditional on $Z \neq 0$ treating the outcome as binary.

The same super learners from the simulation example above were used here in the trial example. The predicted model for the first method is:

$$\Psi_n(W) = -0.01 + 0.02(X_n^{ridge}) + 1.21(X_n^{rf}) - 0.84(X_n^{lars}) - 0.28(X_n^{lm}) + 0.50(X_n^{mars})$$

Where X_n^j is the predicted value for Z based on the j^{th} candidate learner. $j = ridge$ is the ridge regression model. $j = rf$ is the random forests model. $j = lars$ is the least angle regression model. $j = lm$ is the main effects linear

regression model. $j = mars$ is the adaptive regression splines model. The coefficient estimates for each candidate learner from the super learner can be interpreted as a weight for each candidate learner in the final prediction model. Random forests has the largest absolute weight. When interpreting the weights, be cautious of the often near collinearity of the columns of X. To evaluate the super learner in comparison to the candidate learners, a 10-fold cross validation of the super learner and each of the candidate learners themselves was used to estimate $\mathrm{E}(\Psi_n(W) - Z)^2$. Table 4 contains the risk estimates. For the trial example, both the lars algorithm and the

Method	Risk
Lars	0.426
Mars	0.426
Super Learner	0.445
Ridge Regression	0.505
Random Forests	0.509
Linear Model	0.525

Table 4: 10-fold honest cross validation estimates of $\mathrm{E}(\Psi_n(W) - Z)^2$ for the super learner and each of the candidate learners on their own.

mars algorithm outperform the super learner. As observed in the simulation, minimizing the risk $\mathrm{E}(\Psi_n(W) - Z)^2$ should directly relate to minimizing the risk $\mathrm{E}(\Psi_n(W) - \Psi(W))^2$. These cross-validation estimates may be used to select an optimal final model for the treatment decisions.

The second method evaluates $\mathrm{E}(Z|Z \neq 0, W) = m'(W|\beta)$. The estimated super learner model for the second method is:

$$\Psi'_n(W) = -0.53 - 0.40(X_n^{poly}) + 0.55(X_n^{plr}) + 0.81(X_n^{glm})$$

Where X_n^j is the predicted value for Z based on the j^{th} candidate learner. $j = poly$ is the polyclass adaptive spline model. $j = plr$ is the penalized logistic regression model. $j = glm$ is the main effects logistic regression model. To compare the two methods, we created a confidence interval at the mean

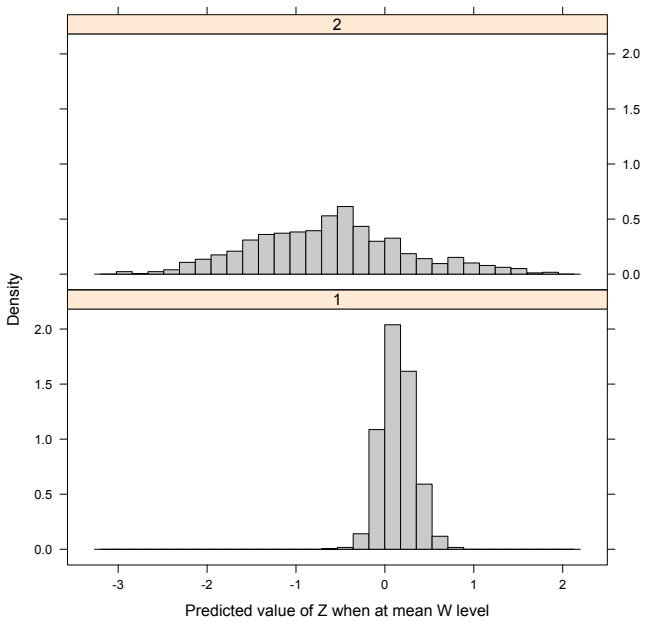


Figure 2: Histograms from 1000 bootstrap samples for $\Psi'(W = \bar{w})$ and $\Psi(W = \bar{w})$. The number in the title bar refers to the method used.

vector for W. Let $\bar{w}$ be the vector of observed means for the baseline variables using all observations in the trial. Confidence intervals were created based on 1000 bootstrap samples of the entire super learner. The 95% confidence interval for $m(\bar{w}|\beta)$ based on the first method is $(-0.20, 0.52)$. The 95% for $m(\bar{w}|\beta)$ based on the second method is $(-2.23, 1.23)$. Although the second method is able to use the distributional information, the penalty for the

smaller sample size is great (308 patients for the first method down to 130 patients for the second method). As can be seen in figure 19.2, the second method has a wide confidence interval compared to the first method.

7 Variable Importance Measure

An additional feature of having a good prediction model is better variable importance measures. The variables in $\mathrm{E}(Z|W)$ are effect modifications and when applying the targeted maximum likelihood estimation (tMLE) variable importance measure [10] the results will be causal effect modification importance measures. The targeted maximum likelihood effect modification variable importance allows the researcher to focus on each variable in W individually while adjust for the other variables in W. An initial variable importance estimate is based on an univariate regression, $Z^* = \beta_{0j} + \beta_{1j} W_j,\ j = 1, \ldots, p$ where p is the number of baseline covariates in W. The top 5 baseline variables based on the ranks of the univariate p-values is presented in table 5. The top unadjusted effect modification variable is an indicator of whether the patients lives in the US or Europe, followed by an indicator for the patients being in RPA class 2, an indicator for the primary tumor being controlled, an indicator for extracranial metastasis, and finally an indicator for the patient's age greater than 65 years. The top 5 baseline variables from the LARS procedure are similar to those from the univariate regression with the exception of Squamous cell indicator replacing the RPA class 2 indicator. For the tMLE variable importance, the effect of W_j on Z is adjusted by all other covariates in W. Let $W_{(-j)}$ be all covariates in W excluding the j^{th} variable.

The targeted maximum likelihood variable importance measure as outlined in [11] was then applied using the predictions from the super learner as the initial estimate of $\mathrm{E}(Z|W)$. The targeted effect modification parameter is then:

$$\psi_j = \mathrm{E}\left(\mathrm{E}(Z|W_j = 1, W_{(-j)}) - \mathrm{E}(Z|W_j = 0, W_{(-j)})\right),\ j = 1, \ldots, p \qquad (8)$$

The top 5 baseline variables are presented in table 5. The effect estimates from the tMLE procedure can be considered causal effect modifiers. Only extracranial mets appears in both the adjusted and unadjusted top 5 list, although squamous cell indicator does appear in both the LARS procedure and the tMLE procedure. The top variable (Mets Dx $>$ 6 Mo) is an indicator for the metastasis diagnosis occurring greater than 6 months after prvious cancer. The tMLE list contains two indicators for histology of the tumor cells (Squamous and Adeno carcinoma) suggesting that some tumor types my respond better to the treatment compared to others. Comparing the variable importance lists, the indicator for the patient being in the United States compared to Europe is on top of the list for the univariate regression and the lars model, but absent from the tMLE list. There is no biological evidence for geographical location to interact with the treatment in this trial. The variable importance based on targeted maximum likelihood is able to appropriately adjust for the confounding on the other variables in W and remove the US versus Europe indicator from the list of top variables. The variable importance list from the tMLE has a better interpretation and is informative as to which patient characteristics have a causal interaction with

Method	Baseline Variable	Effect	p-value
Univariate Regression	US vs Europe	-0.222	0.007
	RPA class 2	-0.229	0.017
	Primary tumor control	0.165	0.052
	Extracranial mets	-0.133	0.069
	Age $>$ 65 years	-0.157	0.075
LARS	US vs Europe	-0.124	0.350
	Primary tumor control	0.080	0.405
	Age $>$ 65 years	-0.050	0.412
	Extracranial mets	-0.028	0.413
	Squamous cell	0.034	0.419
tMLE	Mets Dx $>$ 6 Mo	0.864	$<$0.001
	Squamous cell	1.012	$<$0.001
	Adeno carcinoma	0.129	0.007
	Extracranial mets	-0.102	0.022
	Caucasian	0.172	0.035

Table 5: Top 5 effect modifiers based on univariate regression, lars, and super learner with targeted maximum likelihood. The standard error was based on a bootstrap with 1,000 samples.

the treatment.

8 Discussion

Two methods were proposed for predicting the optimal treatment based on baseline factors. The first method involves modeling Z on W disregarding the knowledge that $\mathrm{E}(Z|W)$ is bounded between -1 and $+1$. The second method incorporates the bounds, but does so at a cost in sample size by modeling $\mathrm{E}(Z|Z \neq 0, W)$. The second method predicts a scaled version of the parameter of interest, and so is still valid for making treatment decisions. In the simulation and trial example presented here, the loss of sample size in the second method greatly increased the variability of the final prediction. But both the simulation and trial example had a high fraction of patients with $Z = 0$ (equivalently, $Y = 0$). The second method may outperform the first method in settings where $\Pr(Y = 0)$ is very small. For the examples presented here, no problems were observed with the first method not respecting the bounds on $\mathrm{E}(Z|W)$.

In the trial example, the super learner did perform better than the main terms linear regression based on the estimate of the risk $\mathrm{E}(\Psi_n(W) - Z)^2$. Even though the super learner has shown to have excellent performance across a range of simulations [2, 12] and in various of our data analyses in breast cancer research, there is a risk that the super learner will result in a slight over-fit. In the data analysis we observed that the super learner was ranked third, but competitive with the top two candidate learners, LARS and MARS. We have also proposed an extension to the super learner outlined

here to adaptively select the number of candidates [13] so that the weaker candidates are not selected, which we believe will protect the super learner against possible over-fitting, but this was not implemented in the current data analysis yet.

We observed that the difference in sample size between the two methods may make the second method unusable in this example, but the two methods also differed in the treatment of right censoring. The first method incorporated the doubly robust censoring unbiased transformation while the second method used the inverse probability of censoring weights. If the model for $Q(W)$ was correctly specified, but the model for the censoring mechanism was not consistently estimating $\pi(W)$, the doubly robust estimator would still be unbiased but the inverse probability of censoring weighted method will be biased. Alternatively, if $\pi(W)$ was correctly specified, but $Q(W)$ was inconsistent, then both methods will be unbiased. The doubly robust transformation gives the researcher two chances to correctly the nuisance parameters, while the inverse weighting method relies solely on the model for $\pi(W)$. When there is uncertainty regarding the model for the censoring mechanism, the doubly robust transformation is preferred.

The methods presented above are not limited to randomized clinical trials. Optimal treatment prediction models could also be estimated from observational or registry data sets. As long as the variables needed to estimate $\Pr(A = 1|W))$ are collected in the study the above methods easily extend to the non-randomized setting. Registry data sets are often larger than randomized trials and therefore have more power to detect the interaction effects necessary for predict optimal treatments.

References

[1] M. J. van der Laan. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2, 2006.

[2] M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(25), 2007.

[3] D. Rubin and M. J. van der Laan. Doubly robust censoring unbiased transformations. Technical Report 208, University of California, Berkeley, Division of Biostatistics, 2006.

[4] A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(3):55–67, 1970.

[5] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[6] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[7] J. H. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19(1):1–141, 1991.

[8] C. Kooperberg, S. Bose, and C. J. Stone. Polychotomous regression. *Journal of the American Statistical Association*, 92:117–127, 1997.

[9] M. Y. Park and T. Hastie. l_1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society, Series B*, 69(4):659–677, 2007.

[10] M. J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *International Journal of Biostatistics*, 2(1), 2007.

[11] O. Bembom, M. L. Petersen, S. Rhee, W. J. Fessel, S. E. Sinisi, R. W. Shafer, and M. J. van der Laan. Biomarker discovery using targeted maximum likelihood estimation: Application to the treatment of antiretroviral resistant HIV infection. Technical Report 221, University of California, Berkeley, Division of Biostatistics, 2007.

[12] S. E. Sinisi, E. C. Polley, , M. L. Petersen, S.Y. Rhee, and M. J. van der Laan. Super learning: An application to the prediction of HIV-1 drug resistance. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007.

[13] E. C. Polley and M. J. van der Laan. Adaptive selection of the functional form for the super learner. in preparation, 2008.

5.3 *Simple, Efficient Estimators of Treatment Effects in Randomized Trials Using Generalized Linear Models to Leverage Baseline Variables*

The following work is currently unpublished elsewhere.

Simple, Efficient Estimators of Treatment Effects in Randomized Trials Using Generalized Linear Models to Leverage Baseline Variables

Michael Rosenblum and Mark J. van der Laan

August 26, 2009

Models, such as logistic regression and Poisson regression models, are often used to estimate treatment effects in randomized trials. These models leverage information in variables collected before randomization, in order to obtain more precise estimates of treatment effects. However, there is the danger that model misspecification will lead to bias. We show that certain easy to compute, model-based estimators are asymptotically unbiased regardless of whether the model used is misspecified or not. Furthermore, these estimators are locally efficient. As a special case of our main result, we consider a simple Poisson model containing only main terms; in this case, the maximum likelihood estimate of the coefficient corresponding to the treatment variable is an asymptotically unbiased estimator of the log rate ratio, even when the model is misspecified. Our results demonstrate one application of targeted maximum likelihood estimation.

1 Introduction

The appropriate use of models in analyzing the results of randomized trials has been the focus of many recent papers (e.g. Pocock et al. (2002); Rosenbaum (2002); Tsiatis et al. (2007); Moore and van der Laan (2007); Freedman (2007a,b, 2008); Zhang et al. (2008); Rosenblum and van der Laan (2008)). Here we present a large class of simple to compute, model-based estimators of treatment effects–the same effects estimated by the intention to treat estimator. Our estimators are asymptotically unbiased, and leverage baseline variables to try to get more precision than the intention to treat estimator (though in some cases it is possible that the intention to treat estimator has greater precision). All of our results hold even when the models used are misspecified, that is, when the models used do not contain the data generating distribution. This is an important property since in practice, models will often be misspecified. Our results demonstrate an application of targeted maximum likelihood estimation, a general estimation method with broad applicability to randomized trials and observational studies described in van der Laan and Rubin (2006).

In the next section, we describe the estimation problem being considered and present related work. Then, in Section 3 we give a brief overview of targeted maximum likelihood methodology, of which our estimators are one application. Our class of estimators and our main result are presented in Section 4. We show how to construct confidence intervals and compute p-values in Section 5. Proofs of our results are given in the Appendix.

2 Description of Estimation Problem, Assumptions, and Related Work

We consider a randomized trial with n subjects, in which a set of baseline variables, denoted by V, are measured. After these variables are measured, subjects are randomized with probability 1/2 to either the treatment or control arm, independent of the baseline variables. We let A denote the treatment assignment, with $A = 1$ corresponding to the treatment arm and $A = 0$ corresponding to the control arm. We denote the outcome variable by Y. For each subject i, we denote their data by the vector (V_i, A_i, Y_i), representing their baseline measurements, treatment assignment, and outcome, respectively. In general, we recommend that baseline variables that are highly predictive of the outcome should be included in the vector V.

Assumptions on the Data Generating Distribution:
We assume that the observations (V_i, A_i, Y_i) are independent, identically distributed draws from an unknown data generating distribution.[1] We also assume the values of all variables are bounded. We assume that A and V are independent, which is ensured by randomization.

Assumptions on the Form of the Generalized Linear Model:
We assume that we are using a generalized linear model with canonical link function and with linear part containing A as a main term and containing an intercept. We assume the exponential family used is one of the following commonly used families: Normal, Binomial, Poisson, Gamma, or Inverse Normal (see McCullagh and Nelder (1998) for definitions of these exponential families). We denote the linear part of the generalized linear model by $\eta = \sum_{i=1}^{k} \beta_j f_j(A, V)$, where $f_1(A, V) = 1$, $f_2(A, V) = A$, and f_j are functions of A and V that are bounded on compact subsets of $\{0, 1\} \times \mathbf{R}^d$, where V is a d-dimensional vector of baseline variables. We also assume the terms $f_j(A, V)$ are linearly independent. (Linear dependencies will be detected by standard statistical software.) Also, we assume that there exists a maximizer β^* of the expected log likelihood that has components with absolute values smaller than some pre-specified bound M. This can be detected, for large enough sample size n, as described in the Appendix.

For the Gamma and Inverse Normal families, where the outcome variable is assumed to take values in $(0, \infty)$, we assume that f_j take positive values and are bounded away from 0 by some $\delta > 0$; also for these two families we restrict β_j to take positive values and be bounded away from 0 by some $\delta > 0$. Furthermore,

[1] This assumption is not guaranteed by randomization. For discussion of this issue, see (Rosenbaum, 2002; Freedman, 2008; Rosenblum and van der Laan, 2008).

for these families, we assume that there exists a maximizer β^* of the expected log likelihood for which all components of β^* are strictly greater than δ.

The above class of generalized linear models includes (but is not limited to) the following examples:

1. Least Squares Regression: For Y continuous, the Normal model assuming $E(Y|A,V)$ has the form:
$$\mu_1(A,V|\beta) = \beta_0 + \beta_1 A + \beta_2 V + \beta_3 AV + \beta_4 V^2,$$
2. Logistic Regression: For Y binary and $\text{logit}(x) = \log(x/(1-x))$, the following model for $P(Y=1|A,V)$:
$$\mu_2(A,V|\beta) = \text{ logit}^{-1}\left(\beta_0 + \beta_1 A + \beta_2 V\right),$$
3. Poisson Regression: For Y a "count" (that is, Y a nonnegative integer), the Poisson (log-linear) model with mean of Y given A, V of the form:
$$\mu_3(A,V|\beta) = \exp\left(\beta_0 + \beta_1 A + \beta_2 V + \beta_3 AV\right).$$
4. Gamma Regression: For Y positive, real valued, the Gamma model with mean of Y given A, V modeled by:
$$\mu_4(A,V|\beta) = 1/\left(\beta_0 + \beta_1(1+A) + \beta_2\exp(V) + \beta_3\exp(AV)\right),$$
where all coefficients β_j are assumed to be positive and bounded away from 0 by some $\delta > 0$.
5. Inverse Normal Regression: For Y positive, real valued, the Inverse Normal model with mean of Y given A, V modeled by:
$$\mu_5(A,V|\beta) = 1/\sqrt{\beta_0 + \beta_1(1+A) + \beta_2\exp(V)},$$
where all coefficients β_j are assumed to be positive and bounded away from 0 by some $\delta > 0$.

We consider estimation and inference for parameters that are smooth functions of the mean effects of being assigned to the two study arms: $E(Y|A=0)$ and $E(Y|A=1)$. This class of parameters includes the difference in means $E(Y|A=1) - E(Y|A=0)$, the ratio of means (or rate ratio) $E(Y|A=1)/E(Y|A=0)$, and the log odds ratio $\log \frac{P(Y=1|A=1)/(1-P(Y=1|A=1))}{P(Y=1|A=0)/(1-P(Y=1|A=0))}$, for example. (Throughout the paper "log" refers to the natural logarithm.) The intention to treat estimator estimates these parameters by substituting the sample means in the control arm and treatment arm, respectively, for $E(Y|A=0)$ and $E(Y|A=1)$. We will denote $E(Y|A=0)$ and $E(Y|A=1)$ by E_0 and E_1, respectively.

Moore and van der Laan (2007) applied targeted maximum likelihood methodology to prove that certain easy to compute estimators based on a logistic regression model are asymptotically unbiased (and locally efficient) even when the model used is misspecified. Our results generalize this important result to a larger class of generalized linear models that includes Normal (Gaussian) models, Poisson models with log link, and models based on the Gamma distribution (with reciprocal link) and Inverse Gaussian distribution (with link $1/\mu^2$). We note that estimation of the risk difference using a Normal model with only main terms corresponds to ANCOVA (analysis of covariance), which has been shown to be asymptotically unbiased even when the model is misspecified (Freedman, 2007a). We also note that in the special case of logistic regression, Freedman (2008) proved a related result under the framework of randomization inference.

Our result for the special case of a Poisson model with only main terms (see the Corollary in Section 4) is a generalization of a result of Gail (1986) that required much stronger assumptions than used here.

Robinson and Jewell (1991) compare the precision of estimators of the marginal effect and estimators of the conditional effect of a treatment, based on linear and logistic regression models. In this paper we focus only on estimating marginal effects, that is, comparisons of $E(Y|A=1)$ and $E(Y|A=0)$. These are the same quantities estimated by the intention to treat estimator. Our estimators leverage baseline variables in order to try to get more precise (i.e. smaller asymptotic variance) estimates than the intention to treat estimator. We note that whether marginal effects or conditional effects are more relevant will depend on the application at hand. Though the focus of this paper is estimation and inference, certain results have been shown for hypothesis testing, in which model-based tests have correct Type I error even when models are misspecified (Rosenblum and van der Laan, 2008).

3 Brief Description of Targeted Maximum Likelihood Estimation

Theorem 1 in the next section is proved using targeted maximum likelihood methodology. We give a brief overview here; a full description is given in (van der Laan and Rubin, 2006). Targeted maximum likelihood is a general methodology for estimation and inference. It can be used to estimate finite-dimensional, pathwise differentiable parameters (such as those considered in this paper) as well as more general parameters including infinite-dimensional, non-pathwise differentiable parameters.

Targeted maximum likelihood estimation has several important advantages over standard maximum likelihood estimation and estimating function-based methodologies. When estimating parameters in the nonparametric model[2], maximum likelihood estimation based on assuming a parametric model (or based on selecting a parametric model using a sieve) may suffer severe bias due to model misspecification; targeted maximum likelihood estimation, on the other hand, only models the parameter of interest, thereby reducing bias due to misspecified models of nuisance parameters. Estimating function based methodology (Robins, 1986, 1987; van der Laan and Robins, 2002), which involves only modeling the parameter of interest, still has important limitations. These include (1) in general not having a satisfactory way to deal with multiple solutions to an estimating equation, (2) only applying to problems that can be expressed in terms of a parameter of interest and a variation independent nuisance parameter, and (3) not being invariant to monotone transformations of the parameter of interest. Targeted maximum likelihood does not have any of these limitations. In addition, in many situations, targeted maximum likelihood can be simply implemented using standard statistical software.

We now give a brief overview of the general algorithm for constructing the targeted maximum likelihood estimator. For a given parameter of interest ψ, the targeted maximum likelihood estimator is constructed in the following six steps: (We also give an oversimplified example, to illustrate these steps here; in the Appendix we go through the same six steps below, in estimating the more general parameters covered in Theorem 1 in the next section.)

1. An initial estimate p_0 of the density of the data generating distribution is constructed, by any method. For example, standard maximum likelihood estimation using a parametric model could be used to generate p_0.

2. The efficient influence curve for the parameter ψ in the nonparametric model is computed, at p_0. Methods for finding the efficient influence curve for a wide variety of parameters can be found in (van der Laan and Robins, 2002). As an example, the efficient influence curve for the mean of random variable Y in the nonparametric model is $Y - \psi$; at a given density p, the efficient influence curve would then be $Y - E_p(Y)$, where E_p is the expectation with respect to the density p.

3. A parametric model with parameter ϵ and corresponding densities $\{p(\epsilon)\}$ is constructed that (i) equals the initial density p_0 at $\epsilon = 0$ and (ii) has score at $\epsilon = 0$ whose linear span contains the efficient influence curve in the nonparametric model, at p_0. Continuing with our simple example, if the parameter is the mean of a continuous random variable Y, and the initial density p_0 was chosen to be a normal distribution with mean $\hat{\mu}$ equal to the sample mean and variance $\hat{\sigma}^2$ equal to the sample variance, then one could choose as parametric model a normal model with the same variance, but with mean equal to $\hat{\mu} + \epsilon\hat{\sigma}^2$. As required, (i) the parametric model at $\epsilon = 0$ is p_0, and (ii) the score at $\epsilon = 0$ is $Y - E_{p_0}(Y)$, which equals the efficient influence curve given in the previous step.

4. The parameter ϵ of the parametric model from the previous step is estimated using maximum likelihood estimation. The new density p_1 is then set to be the density corresponding to $p(\hat{\epsilon})$, where $\hat{\epsilon}$ is the maximum likelihood estimate of ϵ. Continuing our example, since the model is a normal model, this corresponds to estimating ϵ using ordinary least squares regression, and then setting p_1 to be the

[2]By nonparametric model, we generally mean the model consisting of all continuous densities with respect to a given dominating measure. In this paper, we also use"nonparametric model" to describe the model that makes no assumptions on the density of the data generating distribution except that treatment A is randomized, so is independent of baseline variables V.

normal density with mean $\hat{\mu} + \hat{\epsilon}\hat{\sigma}^2$ and variance $\hat{\sigma}^2$. Note that the values $\hat{\mu}$ and $\hat{\sigma}^2$ are considered fixed values when the maximum likelihood estimate for ϵ is computed.

5. We then replace the initial density estimate p_0 by our new density p_1, and repeat steps 2-4 until the algorithm converges to a final density p. In many cases, such as those considered in this paper, the algorithm will have converged (that is, $\hat{\epsilon} = 0$) after steps 2-4, and so no iterations are required (and we say the algorithm converged in 0 steps). In the example of a single random variable Y and parameter the mean of Y, we have such convergence immediately after steps 2-4.

6. Once the algorithm converges to a final density p, the targeted maximum likelihood estimator for the parameter ψ is the plug-in estimator of ψ at p. Continuing our example, where ψ is the mean of Y, the plug-in estimator of ψ at p is the mean of Y under the density p. (See (Bickel and Doksum, 2001, Section 2.1.2) for definition and discussion of the plug-in estimator.)

4 Main Result

Below we present our class of simple estimators based on generalized linear models that are asymptotically unbiased even when the model used is incorrectly specified. We then give the main result of the paper in Theorem 1. We illustrate the theorem with two examples based on Poisson regression models.

The class of estimators is constructed as follows, for any generalized linear model with canonical link function, and any continuously differentiable function r:

1. Estimate the coefficients $\{\beta_j\}$ in the linear part of the generalized linear model using maximum likelihood estimation.

2. Compute $\hat{E}_0 := \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}(0, V_i)$, and $\hat{E}_1 := \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}(1, V_i)$, where $\hat{\mu}(a, v)$ is the predicted value of Y, based on the fit of the generalized linear model, for study arm assignment a and baseline variables v.[3]

3. Compute $r(\hat{E}_0, \hat{E}_1)$; this is our estimator of the parameter $r(E(Y|A = 0), E(Y|A = 1))$.

4. Confidence intervals can be obtained based on estimates of the efficient influence curve (as described in Section 5) or based on the nonparametric bootstrap.

We have the following theorem stating that the above estimator is asymptotically unbiased and locally efficient. The theorem assumes the existence of a maximizer β^* of the expected log-likelihood (where the expectation is taken with respect to the data generating distribution). As proved in the Appendix, one can detect whether such a maximizer exists, with probability tending to 1 as sample size goes to infinity.

Theorem 1: *Consider any generalized linear model from the Normal, Binomial, Poisson, Gamma, or Inverse Gaussian family, with canonical link function, in which the linear part contains the treatment variable as a main term and also contains an intercept. Let r be any continuously differentiable function. Under the assumptions in Section 2, and assuming a maximizer β^* of the expected log-likelihood exists, the above procedure gives an asymptotically unbiased and locally efficient estimator for the parameter $r(E(Y|A = 0), E(Y|A = 1))$, even when the generalized linear model is misspecified. Furthermore, the confidence intervals constructed in Section 5 have asymptotically correct coverage, even when the model is misspecified.*

The class of estimators in Theorem 1 are derived from targeted maximum likelihood methodology (van der Laan and Rubin, 2006), as described in the Appendix. We point out that in this special case of a randomized trial (the case considered throughout this paper), the particular version of the targeted maximum

[3] $\hat{\mu}$ is formally defined in the Appendix, where we also give R code for computing $\hat{E}_0$ and $\hat{E}_1$.

likelihood estimator given in the Appendix coincides with certain g-computation estimators (Robins, 1986, 1987), doubly-robust estimators (Robins, 2000; Robins and Rotnitzky, 2001; Neugebauer and van der Laan., 2002; van der Laan and Robins, 2002), and estimators in (Tsiatis, 2006; Zhang et al., 2008). In addition, the estimator given in Theorem 1 solves the doubly robust estimating equation, and thereby the theory of statistical inference developed in (van der Laan and Robins, 2002) applies, and could alternatively be used to establish that this estimator is asymptotically unbiased and locally efficient even under model misspecification. In general, targeted maximum likelihood estimators will differ from from g-computation estimators, doubly-robust estimators, and estimators in (Tsiatis, 2006; Zhang et al., 2008). We also point out that Theorem 1 holds under the slightly weaker condition that 1 and A are in the linear span of the terms in the linear part η of the generalized linear model; this is important in applying Theorem 1 to models μ_4 and μ_5 listed in Section 2.

To illustrate the above theorem, consider a Poisson model with log link function, and linear part $\eta = \beta_0 + \beta_1 A + \beta_2 V$. We will estimate the log rate ratio of the treatment compared to the control: $\log(E(Y|A = 1)/E(Y|A = 0))$, using this Poisson model. This corresponds to choosing the function r in the theorem to be $r(x, y) = \log(y/x)$. We follow the steps given above the theorem to compute an estimate of the log rate ratio. First, we use maximum likelihood estimation to produce estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ for the coefficients $\beta_0, \beta_1, \beta_2$. Next, we compute

$$\hat{E}_0 := \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}(0, V_i) = \frac{1}{n}\sum_{i=1}^{n} \exp(\hat{\beta}_0 + \hat{\beta}_2 V_i)$$

and

$$\hat{E}_1 := \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}(1, V_i) = \frac{1}{n}\sum_{i=1}^{n} \exp(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 V_i).$$

Lastly, we compute

$$r(\hat{E}_0, \hat{E}_1) = \log(\hat{E}_1/\hat{E}_0) = \log[\sum_{i=1}^{n} \exp(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 V_i)/\sum_{i=1}^{n} \exp(\hat{\beta}_0 + \hat{\beta}_2 V_i)].$$

In this special case, we see that the above estimator can be simplified, leaving as final estimate $\hat{\beta}_1$, the coefficient of the treatment term. Thus, the above theorem implies the following corollary:

Corollary: *Consider a Poisson model with only main terms A and V (where V is a vector of pre-randomization variables). Under the assumptions in Section 2, and assuming a maximizer β^* of the expected log-likelihood exists, we have $\hat{\beta}_1$, the estimate of the coefficient corresponding to the treatment term A, is an asymptotically unbiased estimate of the log rate ratio, even when the model is misspecified. Also, the confidence intervals constructed in Section 5 have asymptotically correct coverage. Furthermore, this estimator is locally efficient in that when the Poisson model is correctly specified this estimator attains the efficiency bound for the model that only assumes treatment assignment A is independent of baseline variables V.*

As another example, consider the problem of estimating the log rate ratio using a Poisson model with log link function, but this time with the linear part containing an interaction term: $\eta = \beta_0 + \beta_1 A + \beta_2 V + \beta_3 AV$. Again, we use maximum likelihood estimation to get estimates for the coefficients $\beta_0, \beta_1, \beta_2, \beta_3$; as above, we use as estimator $r(\hat{E}_0, \hat{E}_1)$, which equals

$$\begin{aligned}\log(\hat{E}_1/\hat{E}_0) &= \log[\sum_{i=1}^{n} \exp(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 V_i + \hat{\beta}_3 V_i)/\sum_{i=1}^{n} \exp(\hat{\beta}_0 + \hat{\beta}_2 V_i)] \\ &= \hat{\beta}_1 + \log[\sum_{i=1}^{n} \exp((\hat{\beta}_2 + \hat{\beta}_3) V_i)/\sum_{i=1}^{n} \exp(\hat{\beta}_2 V_i)].\end{aligned}$$

The proof of Theorem 1, given in the Appendix, applies the targeted maximum likelihood algorithm to the application in this paper, namely, estimating a function r of the conditional means given assignment to the treatment arm and the control arm, respectively. It turns out in this case, that when the initial density p_0 is chosen based on the maximum likelihood estimate using a generalized linear model for Y given A, V, and a canonical link is used, then the targeted maximum likelihood algorithm converges in zero steps (as defined in Section 3) and has the simple form given just before Theorem 1. The reason is that the score of a generalized linear model with canonical link function has a simple form that is closely related to the efficient influence curve of the conditional means $E(Y|A = 0)$ and $E(Y|A = 1)$ in the model that is nonparametric except for assuming A is randomized (that is, independent of baseline variables V).

We point out that the choice of generalized linear model, including selecting which terms to include in the linear part, should be pre-specified in the study protocol, to avoid possible data snooping in choosing a model.

5 Statistical Inference: Computing Confidence Intervals and p-values

We show how to compute confidence intervals and p-values for the estimator given just before Theorem 1. We use the method from Section 4 of (Moore and van der Laan, 2007), based on estimates of the efficient influence curve of our parameter in the nonparametric model. This involves first computing an estimate $\hat{\sigma}^2$ for the asymptotic variance of $\sqrt{n}(\hat{\psi} - \psi)$, where $\hat{\psi} = r(\hat{E}_0, \hat{E}_1)$ is our estimator and ψ is the (unknown) value for the parameter $r(E_0, E_1)$ that we are estimating; we describe how to compute $\hat{\sigma}^2$ below. Having computed $\hat{\sigma}^2$, we next compute an 0.95 confidence interval $(\hat{\psi} - 1.96\hat{\sigma}/\sqrt{n}, \hat{\psi} + 1.96\hat{\sigma}/\sqrt{n})$. Also, we can test the null hypothesis $\psi = \psi_0$ using the test statistic $T = \sqrt{n}(\hat{\psi} - \psi_0)/\hat{\sigma}$, which is asymptotically normally distributed with mean 0 and variance 1 under this null hypothesis and under the regularity conditions given in Section 2. The confidence interval and p-value computed by this method are asymptotically correct, even when the generalized linear model used is incorrectly specified.

The above procedures rely on an estimate of the asymptotic variance of $\sqrt{n}(\hat{\psi} - \psi)$, which we denote by $\hat{\sigma}^2$, and define now. It can be computed based on the partial derivatives of the function r used in defining our parameter and on estimates of the efficient influence curve of (E_0, E_1) in the nonparametric model. Let r'_1, r'_2 denote the partial derivatives of the function r with respect to the first component and second component, respectively. For example, when our parameter is the rate ratio, then $r(x, y) = y/x$, and so $r'_1(E_0, E_1) = -E_1/E_0^2, r'_2(E_0, E_1) = 1/E_0$. Define the vector with two components:

$$\begin{aligned} D(p)(V, A, Y) &= (D_1(p)(V, A, Y), D_2(p)(V, A, Y)) \\ &= ((1 - A)(Y - E_p(Y|A = 0, V))/p(A = 0) + E_p(Y|A = 0, V) - E_p(Y|A = 0), \\ &\quad A(Y - E_p(Y|A = 1, V))/p(A = 1) + E_p(Y|A = 1, V) - E_p(Y|A = 1)) \,, \end{aligned} \tag{1}$$

where E_p is the expectation with respect to the density p. The efficient influence curve for (E_0, E_1) in the nonparametric model is $D(p^*)$, where the density p^* is that of the (unknown) data generating distribution. (See van der Laan and Robins (2002) for the derivation of this efficient influence curve.)

As in Theorem 1, assume there exists a maximizer β^* of the expected log likelihood of the generalized linear model, where the expectation is with respect to the (unknown) data generating distribution. Under this assumption, we show in the Appendix that such a maximizer is unique, and that the maximum likelihood estimator $\hat{\beta}$ converges to β^*. Let $p(\beta^*)$ be the density of Y given A, V corresponding to the parameter $\beta = \beta^*$ in the generalized linear model. In terms of D and r'_1, r'_2, the asymptotic variance of $\sqrt{n}(\hat{\psi} - \psi)$ is

$$\sigma^2 = E_{p^*} \left(r'_1(E_0, E_1) D_1(p(\beta^*)(V, A, Y) + r'_2(E_0, E_1) D_2(p(\beta^*)(V, A, Y)) \right)^2 . \tag{2}$$

We estimate this by

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \left(r'_1(\hat{E}_0, \hat{E}_1) D_1(\hat{p})(V_i, A_i, Y_i) + r'_2(\hat{E}_0, \hat{E}_1) D_2(\hat{p})(V_i, A_i, Y_i) \right)^2 ,$$

where $\hat{p}$ is the density estimated by targeted maximum likelihood given in the Appendix. Since as shown in the Appendix, $E_{\hat{p}}(Y|A=0,V)=\hat{\mu}(0,V)$ and $E_{\hat{p}}(Y|A=1,V)=\hat{\mu}(1,V)$, where $\hat{\mu}(a,v)$ is the predicted mean of Y given $A=a,V=v$ based on the maximum likelihood estimate for the generalized linear model, we have

$$\begin{aligned}\hat{\sigma}^2 &= \sum_{i=1}^{n}\Big(r_1'(\hat{E}_0,\hat{E}_1)[(1-A_i)(Y_i-\hat{\mu}(0,V_i))/(1/2)+\hat{\mu}(0,V_i)-\hat{E}_0] \\ &\quad + r_2'(\hat{E}_0,\hat{E}_1)[A_i(Y_i-\hat{\mu}(1,V_i))/(1/2)+\hat{\mu}(1,V_i)-\hat{E}_1]\Big)^2.\end{aligned}$$

For example, when our parameter is the rate ratio, so that as argued above $r_1'(E_0,E_1)=-E_1/E_0^2, r_2'(E_0,E_1)=1/E_0$, we have

$$\begin{aligned}\hat{\sigma}^2 &= \sum_{i=1}^{n}\Big(-\hat{E}_1/\hat{E}_0^2[(1-A_i)(Y_i-\hat{\mu}(0,V_i))/(1/2)+\hat{\mu}(0,V_i)-\hat{E}_0] \\ &\quad + 1/\hat{E}_0[A_i(Y_i-\hat{\mu}(1,V_i))/(1/2)+\hat{\mu}(1,V_i)-\hat{E}_1]\Big)^2.\end{aligned}$$

Having now computed $\hat{\sigma}^2$, one can use this in the formulas given in the first paragraph of this section to compute confidence intervals and p-values.

6 Appendix: Proof of Theorem 1

We prove Theorem 1. Consider the model used throughout this paper, where the data consist of i.i.d. observations (V_i,A_i,Y_i) and the randomized treatment A_i is assumed to take values 0 and 1 with probability 1/2, independent of the baseline variables V_i. The parameter being estimated is a smooth function r of the conditional means $E(Y|A=0)$ and $E(Y|A=1)$. The efficient influence curve for this parameter in the nonparametric model is then a linear combination of the efficient influence curves for the conditional means $E(Y|A=0)$ and $E(Y|A=1)$. At any given density p, these efficient influence curves are given by

$$D_1(p)(V,A,Y)=(1-A)(Y-E_p(Y|A=0,V))/p(A=0)+E_p(Y|A=0,V)-E_p(Y|A=0), \tag{3}$$

and

$$D_2(p)(V,A,Y)=A(Y-E_p(Y|A=1,V))/p(A=1)+E_p(Y|A=1,V)-E_p(Y|A=1) \tag{4}$$

respectively, where E_p is the expectation with respect to the density p.

We will use a generalized linear model with canonical link. As described in (McCullagh and Nelder, 1998), the density of such a generalized linear model can be represented, for suitable choices of functions b,c as

$$\exp(Y\eta-b(\eta)+c(Y,\phi)), \tag{5}$$

where $\eta=\sum_j\beta_j f_j(A,V)$ is the "linear part" of the model, with terms $f_j(A,V)$ and coefficients β_j, and ϕ is a dispersion parameter.[4] The canonical link function g is defined as $\dot{b}^{-1}$, the inverse of the derivative of the function b. We let $\mu(A,V)$ denote the mean of Y given A,V according to the density (5), where the dependence of $\mu(A,V)$ on β is implicit. We note that $\mu(A,V)=\dot{b}(\eta(A,V))$, which is proved in (Bickel and Doksum, 2001). Also, under the assumptions in Section 2 on our families of generalized linear models with canonical links, we have $\ddot{b}(\eta):=\frac{d^2b}{d\eta^2}>0$ for all η.

We first extract some useful information from the fact that $\hat{\beta}$ is the maximum likelihood estimator of the generalized linear model defined above. Let $p_{01}(Y|A,V)$ denote the the maximum likelihood estimate

[4]For binary outcomes, the function $b(\eta)=\log(1+e^\eta)$ and $c(Y,\phi)=0$. For Poisson regression, in which the outcome is a nonnegative integer, $b(\eta)=e^\eta$ and $c(Y,\phi)=-\log Y!$. Note that in both cases, $\ddot{b}(\eta):=\frac{d^2b}{d\eta^2}>0$ for all η.

for the density of Y given A, V, using the above generalized linear model. Under the regularity assumptions made in Section 2, we have that the derivative of the log likelihood at $\hat{\beta}$ must be 0. The derivative of the log of (5) is $(\partial\eta/\partial\beta)(Y - \dot{b}(\eta)) = (\partial\eta/\partial\beta)(Y - E_{p_{01}}(Y|A_i, V_i)))$, based on the fact for generalized linear models that $\mu(A, V) = \dot{b}(\eta(A, V))$. Since we assumed the linear part η of the generalized linear model contains an intercept term and also contains A as a main term[5], this implies

$$\sum_{i=1}^{n}(Y_i - E_{p_{01}}(Y|A_i, V_i)) = 0, \tag{6}$$

and

$$\sum_{i=1}^{n} A_i(Y_i - E_{p_{01}}(Y|A_i, V_i)) = 0. \tag{7}$$

The targeted maximum likelihood algorithm requires an initial density estimator p_0 for the data generating distribution of (V, A, Y). It will be based on the maximum likelihood estimate $\hat{\beta}$ from the generalized linear model and the set of observed baseline variables $\{V_i\}$. We set

$$p_0(V, A, Y) = p_{01}(Y|A, V)p_{02}(A|V)p_{03}(V), \tag{8}$$

where we have

- $p_{01}(Y|A, V)$ is the maximum likelihood estimate for the density of Y given A, V, using the pre-specified generalized linear model,
- $p_{02}(A|V) = 1/2$, to reflect the known randomization probabilities, and
- $p_{03}(V)$ is the empirical distribution of V.

Since $p_{03}(V)$ was chosen to be the empirical distribution of V, and by our choice of $p_{02}(A|V) = 1/2$, we have

$$\sum_{i=1}^{n}(E_{p_0}(Y|A = 1, V_i) - E_{p_0}(Y|A = 1)) = \sum_{i=1}^{n} E_{p_{01}}(Y|A = 1, V_i) - \sum_{i=1}^{n}[(1/n)\sum_{j=1}^{n} E_{p_{01}}(Y|A = 1, V_j))] = 0. \tag{9}$$

Similarly, we have

$$\sum_{i=1}^{n}(E_{p_0}(Y|A = 0, V_i) - E_{p_0}(Y|A = 0)) = \sum_{i=1}^{n} E_{p_{01}}(Y|A = 0, V_i) - \sum_{i=1}^{n}[(1/n)\sum_{j=1}^{n} E_{p_{01}}(Y|A = 0, V_j))] = 0. \tag{10}$$

We now define our parametric model $\{p(\epsilon)\}$ that satisfies conditions (i) and (ii) of step 3 of the targeted maximum likelihood algorithm outlined in Section 3. It will involve adding a term to the linear part of the generalized linear model and also modifying $p_{03}(V)$. We let $p(\epsilon)$ be defined as $p_{01,\epsilon}(Y|A, V)p_{02}(A|V)p_{03,\epsilon}(V)$, where $\epsilon = (\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)$, for densities $p_{01,\epsilon}(Y|A, V)$ and $p_{03,\epsilon}(V)$ defined next. First, $p_{01,\epsilon}(Y|A, V)$ is defined in terms of the generalized linear model as $\exp(Y\eta' - b(\eta') + c(Y, \phi))$, where $\eta' = \hat{\eta} + \epsilon_1 + \epsilon_2 A$, and $\hat{\eta} = \sum_j \hat{\beta}_j f_j(A, V)$. We note that $\dot{b}(\hat{\eta}(A, V)) = E_{p_{01}}(Y|A, V)$, which follows from the fact that for any generalized linear model, $\dot{b}(\eta(A, V))$ is the mean of Y given A, V according to the model at β, which is proved in (Bickel and Doksum, 2001).
Next, we define

$$p_{03,\epsilon}(V) = C_\epsilon \exp(\epsilon_3(E_{p_{01}}(Y|A = 0, V) - E_{p_{01}}(Y|A = 0)) + \epsilon_4(E_{p_{01}}(Y|A = 1, V) - E_{p_{01}}(Y|A = 1)))p_{03}(V),$$

where C_ϵ is chosen so that $p_{03,\epsilon}(v)$ integrates to 1.

[5]In fact, it suffices that 1 and A are each in the linear span of the terms in the linear part η; this is important for applying Theorem 1 to models μ_4, μ_5 listed in Section 2.

Then $p(\epsilon)$ at $\epsilon = 0$ equals the initial density estimator p_0, and the components of the score of $p(\epsilon)$ at $\epsilon = 0$ equal

$$\frac{d}{d\epsilon_1}[\log p(\epsilon)]|_{\epsilon=0} = \frac{d}{d\epsilon_1}[\log p_{01,\epsilon}(Y|A,V)]|_{\epsilon=0} = (Y - \dot{b}(\hat{\eta})) = (Y - E_{p_{01}}(Y|A,V)), \tag{11}$$

$$\frac{d}{d\epsilon_2}[\log p(\epsilon)]|_{\epsilon=0} = \frac{d}{d\epsilon_2}[\log p_{01,\epsilon}(Y|A,V)]|_{\epsilon=0} = A(Y - \dot{b}(\hat{\eta})) = A(Y - E_{p_{01}}(Y|A,V)), \tag{12}$$

$$\frac{d}{d\epsilon_3}[\log p(\epsilon)]|_{\epsilon=0} = \frac{d}{d\epsilon_3}[\log p_{03,\epsilon}(V)]|_{\epsilon=0} = E_{p_{01}}(Y|A=0,V) - E_{p_{01}}(Y|A=0), \tag{13}$$

$$\frac{d}{d\epsilon_4}[\log p(\epsilon)]|_{\epsilon=0} = \frac{d}{d\epsilon_4}[\log p_{03,\epsilon}(V)]|_{\epsilon=0} = E_{p_{01}}(Y|A=1,V) - E_{p_{01}}(Y|A=1). \tag{14}$$

Thus, the efficient influence curves for $E(Y|A=0)$ and for $E(Y|A=1)$, (3) and (4) above, are in the linear span of the score of $p(\epsilon)$ at $\epsilon = 0$; this satisfies requirement (ii) in step 3 of the targeted maximum likelihood procedure given in Section 3.

We now show that the maximum likelihood estimator of ϵ for the model $\{p(\epsilon)\}$, is 0, whenever the conditions of Theorem 1 hold. By our assumption that the expected log-likelihood has a unique maximizer and the other assumptions in Section 2, we have that for sufficiently large n, the log likelihood $\sum_{i=1}^n \log p(\epsilon)(V_i, A_i, Y_i)$ has a unique maximizer. By strict concavity of the log likelihood (as proved in (Rosenblum and van der Laan, 2008, Appendix D) for our families of generalized linear models with canonical links), the maximum likelihood estimator $\hat{\epsilon}$ is the unique value of ϵ for which $d/d\epsilon[\sum_{i=1}^n \log p(\epsilon)(V_i, A_i, Y_i)] = 0$. Equations (6-10) and (11-14) imply $d/d\epsilon[\sum_{i=1}^n \log p(\epsilon)(V_i, A_i, Y_i)] = 0$ at $\epsilon = 0$, and so $\hat{\epsilon} = 0$ is the maximum likelihood estimator for the model $\{p(\epsilon)\}$. Therefore, the targeted maximum likelihood procedure converges in zero steps. Furthermore, since the final density output by the targeted maximum likelihood algorithm is equal to the initial density estimator p_0, we have that the targeted maximum likelihood estimator of the parameter $(E(Y|A=0), E(Y|A=1))$ is exactly as given in Theorem 1.

Theorem 1 requires the existence of a maximizer β^* of the expected log-likelihood $E(Y\eta - b(\eta) + c(Y,\phi))$, where the expectation is with respect to the data generating distribution. Given the assumptions in Section 2, this is sufficient to ensure that the the maximum likelihood estimator $\hat{\beta}_n$ converges to β^* and that $\sqrt{n}(\hat{\beta}_n - \beta^*)$ is asymptotically normal. This follows from the strict concavity of the expected log-likelihood for generalized linear models with canonical links, proved in (Rosenblum and van der Laan, 2008, Appendix D).

So far we have shown that the targeted maximum likelihood estimator in our setting estimator is of the simple form given in Section 4. We now verify that the regularity conditions given in Section 2 are sufficient to prove all the claims in Theorem 1. To this end, we apply Theorem 1 of van der Laan and Rubin (2006), which under conditions that we verify below, gives that the estimator $r(\hat{E}_0, \hat{E}_1)$ is asymptotically unbiased with asymptotic variance as defined in (2) in Section 5, and is locally efficient.

There are five conditions in Theorem 1 of van der Laan and Rubin (2006), which we verify now. We note that we apply Theorem 1 of van der Laan and Rubin (2006) to the parameter (E_0, E_1) and estimator $(\hat{E}_0, \hat{E}_1)$. This then implies the desired results for the parameter $r(E_0, E_1)$ and estimator $r(\hat{E}_0, \hat{E}_1)$. Denote the density p_0 defined above, at sample size n, by p_0^n. To apply Theorem 1 of van der Laan and Rubin (2006), we need to show:

i. The model is convex.

ii. The parameter (E_0, E_1) is linear.

iii. Our estimator $(\hat{E}_0, \hat{E}_1)$ of (E_0, E_1) satisfies

$$(\hat{E}_0, \hat{E}_1) - (E_0, E_1) = \frac{1}{n}\sum_{i=1}^n D(p_0^n)(V_i, A_i, Y_i) - E_{p^*} D(p_0^n)(V, A, Y),$$

where E_{p^*} is the expectation over the variables V, A, Y with respect to the data generating distribution, and where p_0^n is considered fixed.

iv. $D(p_0^n)$ is in a Donsker class with probability tending to 1.

v. $E_{p^*}\left(D_i(p_0^n)(V,A,Y) - D_i(p(\beta^*))(V,A,Y)\right)^2$ converges to 0 in probability, for $i \in \{1,2\}$, where D_1, D_2 are defined in (3) and (4).

Proof of conditions (i)-(v) above:

Condition (i) follows from our model being nonparametric except for assuming, due to randomization, that $p(A|V) = 1/2$.

Condition (ii) follows since for p_1, p_2 two densities in our model, and defining $p_3 = \lambda p_1 + (1-\lambda)p_2$, for $\lambda \in [0,1]$, we have $p_3(Y|A) = p_3(Y,A)/p_3(A) = \int p_3(v,A,Y)dv/(1/2) = \lambda \int p_1(v,A,Y)dv/(1/2) + (1-\lambda)\int p_2(v,A,Y)dv/(1/2) = \lambda p_1(Y|A) + (1-\lambda)p_2(Y|A)$. Thus, the conditional mean of Y given A under p_3 is the convex combination of these conditional means under p_1 and p_2, which proves linearity of the parameter (E_0, E_1) in our model.

Condition (iii) follows since $\frac{1}{n}\sum_{i=1}^n D(p_0^n)(V_i, A_i, Y_i) = 0$ using the definitions (3), (4) and applying (6), (7), (9), and (10), and since

$$\begin{aligned}
E_{p^*} D_1(p_0^n)(V,A,Y) &= E_{p^*}(1-A)(Y - E_{p_0^n}(Y|A=0,V))/p_0^n(A=0) + E_{p_0^n}(Y|A=0,V) - E_{p_0^n}(Y|A=0) \\
&= E_{p^*}(1-A)(Y)/(1/2) - E_{p^*}E_{p_0^n}(Y|A=0,V) + E_{p^*}E_{p_0^n}(Y|A=0,V) - E_{p^*}E_{p_0^n}(Y|A=0) \\
&= E_{p^*}(1-A)(Y)/(1/2) - E_{p^*}E_{p_0^n}(Y|A=0) \\
&= E_{p^*}(Y|A=0) - \hat{E}_0 \\
&= E_0 - \hat{E}_0
\end{aligned}$$

where the second equality follows using the fact that A and V are independent in all of our densities p_0^n, and the second to last equality follows from E_{p^*} being with respect to V, A, Y and treating p_0^n as fixed; a similar derivation shows the analogous statement for $E_{p^*} D_1(p_0^n)(V,A,Y)$.

To show (iv), first let $\mu_\beta(a,v)$ denote the mean of Y given $A = a, V = v$ according to the generalized linear model (5), which depends on β through the linear part $\eta = \sum_j \beta_j f_j(A,V)$. We will show the class of functions $\{\bar{D}_{\alpha_1,\alpha_2,\beta}(v,a,y)\}$ is Donsker, where we define

$$\begin{aligned}
\bar{D}_{\alpha_1,\alpha_2,\beta}(v,a,y) &= ((1-a)(y - \mu_\beta(0,v))/(1/2) + \mu_\beta(0,v) - \alpha_1, \\
&\qquad a(y - \mu_\beta(1,v))/(1/2) + \mu_\beta(1,v) - \alpha_2),
\end{aligned} \tag{15}$$

and we require $|\alpha_1| \le M, |\alpha_2| \le M, \beta \in B$, for B the set of possible β, defined in Section 2,[6] which was selected to ensure our class is over a bounded parameter set; additionally, our assumptions on boundedness of variables and the functions f_j in Section 2, combined with the functional forms of the generalized linear models we are considering, guarantee that the first and second derivatives of $\bar{D}$ with respect to v, a, y are uniformly bounded. We can then apply the result from (van der Vaart, 1998, Example 19.9, page 272), which implies this class of functions is a Donsker class. Since $D(p_0^n)$ are all contained in this class, condition (iv) above is satisfied. We note that Theorem 1 of van der Laan and Rubin (2006) only requires conditions (i) to (iv) in order to prove consistency of the estimator $(\hat{E}_1, \hat{E}_2)$, which we'll use below in proving (v).

Lastly, we show (v) above holds. Let $p(\beta^*)$ denote the density of Y given A, V defined in (5) corresponding to $\beta = \beta^*$, for β^* the maximizer of the expected log-likelihood $E(Y\eta - b(\eta) + c(Y,\phi))$, where the expectation is with respect to the data generating distribution p^*. Note that whenever the model (5) is misspecified, p^* and $p(\beta^*)$ will be different densities. Our theorem still holds in this case, since the only fact we assume about $p(\beta^*)$ is that it satisfies $E_{p(\beta^*)} f_j(A,V)(Y - \mu_{\beta^*}(A,V)) = 0$, for all j, which follows from (5) and β^* being a maximizer of the corresponding expected log-likelihood. Below we let $\hat{\beta}_n$ denote the maximum

[6] In Section 2, we assumed all components of β must have absolute value at most M for some constant M, and for the Gamma and Inverse Normal families, B is further restricted to contain only β for which all components are positive and more than δ for some $\delta > 0$.

likelihood estimator from the generalized linear model fit at sample size n. We then have the following chain of inequalities, where we let $p_{02}(A|V) = 1/2$ and $p^*_{03}(V)$ be the true density of V:

$$\begin{aligned}
& E_{p^*}\left(D_1(p_0^n)(V,A,Y) - D_1(p(\beta^*)p_{02}p^*_{03})(V,A,Y)\right)^2 \\
= & \int \{(1-a)(y-\mu_{\hat\beta_n}(0,v))/(1/2) + \mu_{\hat\beta_n}(0,v) - \hat E_1 \\
& \quad -[(1-a)(y-\mu_{\beta^*}(0,v))/(1/2) + \mu_{\beta^*}(0,v) - E_1]\}^2 p^*(v,a,y)dvdady \\
= & \int \left((1-2a)(\mu_{\beta^*}(0,v) - \mu_{\hat\beta_n}(0,v)) + E_1 - \hat E_1\right)^2 p^*(v,a,y)dvda \\
\leq & \; 2\int \left([(\mu_{\beta^*}(0,v) - \mu_{\hat\beta_n}(0,v))]^2 + [E_1 - \hat E_1]^2\right) p^*(v,a)dvda \qquad (16) \\
= & \; 2\int \left([(\mu_{\beta^*}(0,v) - \mu_{\hat\beta_n}(0,v))]^2 + [E_1 - \hat E_1]^2\right) p^*(v)dv \qquad (17) \\
\leq & \; C_1||\beta^* - \hat\beta_n||^2 + 2[E_1 - \hat E_1]^2. \qquad (18) \\
& \qquad (19)
\end{aligned}$$

where the the first equality follows from definitions; the second equality follows from canceling terms and noting that there is no longer any dependence on y; the inequality (16) follows from the bound $(x+y)^2 \leq 2(x^2+y^2)$ and noting that $1-2a$ is always either 1 or -1; the equality (17) follows from noting that there is no longer dependence on a; and the last line follows from $\mu_\beta(0,v)$ having first derivative uniformly bounded by a constant. The last line of the above display converges to 0 in probability since by our assumptions in Section 2, $\hat\beta_n$ converges to β^* in probability and also the consistency of $\hat E_1$ follows from conditions (i)-(iv) above, as described in Theorem 1 of van der Laan and Rubin (2006). An analogous bound as just derived proves that $E_{p^*}\left(D_2(p_0^n))(V,A,Y) - D_2(p(\beta^*))(V,A,Y)\right)^2$ converges to 0 in probability.

This completes our verification of the conditions (i)-(v) above of Theorem 1 of van der Laan and Rubin (2006), which implies that the estimator $(\hat E_1, \hat E_2)$ converges to $(E(Y|A=0), E(Y|A=1))$, and that $\sqrt{n}[(\hat E_1, \hat E_2) - (E(Y|A=0), E(Y|A=1))]$ is asymptotically normal with variance given by (2), and is locally efficient in that if the generalized linear model is correctly specified, then (2) achieves the efficiency bound for the nonparametric model. This completes the proof of Theorem 1.
□

Since the data generating distribution is unknown, one generally cannot directly check whether there exists a maximizer β^* of the expected log-likelihood $E(Y\eta - b(\eta) + c(Y,\phi))$, where the expectation is taken with respect to the data generating distribution. However, as proved in (Rosenblum and van der Laan, 2008, Appendix D), by strict concavity of the $E(Y\eta - b(\eta) + c(Y,\phi))$, we always have either (1) there is a unique maximizer of the expected log-likelihood or (2) the Euclidean norm of the maximum likelihood estimator grows without bound as sample size goes to infinity. Thus, for large enough sample size n, one will know whether there exists a maximizer β^* of the expected log-likelihood, based on whether $\hat\beta_n$ exceeds a pre-specified (large) threshold.

We now give R code that computes the estimator given just before Theorem 1. The code below corresponds to the specific example of a Poisson model with log link and linear part $\beta_0 + \beta_1 A + \beta_2 V + \beta_3 AV$.

```
# Given vectors V, A, Y of length n containing baseline variables, treatment assignment
# and outcome, respectively, compute the estimated log rate ratio
modelfit <- glm(Y ~ 1 + A + V+ A*V,family=poisson)
E_0_hat <- mean(predict.glm(modelfit, type = "response", newdata=data.frame(A=rep(0,n),V=V)))
E_1_hat <- mean(predict.glm(modelfit, type = "response", newdata=data.frame(A=rep(1,n),V=V)))
log_rate_ratio_estimate <- log(E_1_hat/E_0_hat)
```

References

Bickel, P. J. and Doksum, K. A. (2001). *Mathematical Statistics*, volume 1. Prentice Hall, Upper Saddle River, New Jersey.

Freedman, D. A. (2007a). On regression adjustments to experimental data. *Advances in Applied Mathematics (To Appear)* .

Freedman, D. A. (2007b). On regression adjustments to experiments with several treatments. *Annals of Applied Statistics (To Appear)* .

Freedman, D. A. (2008). Randomization does not justify logistic regression. *Statistical Science* **23,** 237–249.

Gail, M. H. (1986). Adjusting for covariates that have the same distribution in exposed and unexposed cohorts. *In Modern Statistical Methods in Chronic Disease Epidemiology, Eds. S.H. Moolvankar and R. L. Prentice, New York, Wiley.* pages 3–18.

McCullagh, P. and Nelder, J. A. (1998). *Generalized Linear Models*. Chapman and Hall/CRC, Monographs on Statistics and Applied Probability 37, Boca Raton, Florida, 2nd edition.

Moore, K. L. and van der Laan, M. J. (2007). Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 215. http://www.bepress.com/ucbbiostat/paper215* .

Neugebauer, R. and van der Laan., M. J. (2002). Why prefer double robust estimates? illustration with causal point treatment studies. working paper 115. http://www.bepress.com/ucbbiostat/paper115. *U.C. Berkeley Division of Biostatistics Working Paper Series* .

Pocock, S. J., Assmann, S., Enos, L., and Kasten, L. (2002). Subgroup analysis, covariate adjustment, and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine* **21,** 2917–2930.

Robins, J. (1986). A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect. (with errata). *Mathematical Modelling* **7,** 1393–1512.

Robins, J. M. (1987). A graphical approach to the identication and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Disease. Supplement 2.* **40,** 139–161.

Robins, J. M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science 1999.* pages 6–10.

Robins, J. M. and Rotnitzky, A. (2001). Comment on the bickel and kwon article, "inference for semiparametric models: Some questions and an answer". *Statistica Sinica* **11,** 920–936.

Robinson, L. D. and Jewell, N. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review* **59,** 227–240.

Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statistical Science* **17,** 286–327.

Rosenblum, M. and van der Laan, M. (2008). Using regression to analyze randomized trials: Valid hypothesis tests despite incorrectly specified models. *Biometrics (To Appear). Technical Report available at U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 219. http://www.bepress.com/ucbbiostat/paper219 and Web Appendix at http://people.csail.mit.edu/mrosenblum/regressionappendix.pdf* .

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer Science and Business Media, LLC.

Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. (2007). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine (To Appear)* .

van der Laan, M. J. and Robins, J. M. (2002). *Unified methods for censored longitudinal data and causality*. Springer, New York.

van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 213. http://www.bepress.com/ucbbiostat/paper213* .

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, New York.

Zhang, M., Tsiatis, A. A., and Davidian, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics* **64,** 707–715.

Chapter 6

Realistic Individualized Treatment Rules in Observational Studies

6.1 *Estimating the Effect of Vigorous Physical Activity on Mortality in the Elderly Based on Realistic Individualized Treatment and Intention-to-Treat Rules*

The following article appears as it was originally published on the University of California, Berkeley Division of Biostatistics Working Paper Series website in 2007, http://www.bepress.com/ucbbiostat/paper217/.

It was later published as *A Practical Illustration of the Importance of Realistic Individualized Treatment Rules in Causal Inference* in *Electronic Journal of Statistics* in 2007, http://works.bepress.com/bembom/21/.

Estimating the effect of vigorous physical activity on mortality in the elderly based on realistic individualized treatment and intention-to-treat rules

Oliver Bembom and Mark J. van der Laan

Division of Biostatistics, University of California at Berkeley

Abstract

The effect of vigorous physical activity on mortality in the elderly is difficult to estimate using conventional approaches to causal inference that define this effect by comparing the mortality risks corresponding to hypothetical scenarios in which all subjects in the target population engage in a given level of vigorous physical activity. A causal effect defined on the basis of such a static treatment intervention can only be identified from observed data if all subjects in the target population have a positive probability of selecting each of the candidate treatment options, an assumption that is highly unrealistic in this case since subjects with serious health problems will not be able to engage in higher levels of vigorous physical activity. This problem can be addressed by focusing instead on causal effects that are defined on the basis of realistic individualized treatment rules and intention-to-treat rules that explicitly take into account the set of treatment options that are available to each subject. We present a data analysis to illustrate that estimators of static causal effects in fact tend to overestimate the beneficial impact of high levels of vigorous physical activity while corresponding estimators based on realistic individualized treatment rules and intention-to-treat rules can yield unbiased estimates. We emphasize that the problems encountered in estimating static causal effects are not restricted to the IPTW estimator, but are also observed with the G-computation estimator, the DR-IPTW estimator, and the targeted MLE. Our analyses based on realistic individualized treatment rules and intention-to-treat rules suggest that high levels of vigorous physical activity may confer reductions in mortality risk on the order of 15-30%, although in most cases the evidence for such an effect does not quite reach the 0.05 level of significance.

1 Introduction

A substantial body of epidemiologic research indicates that recent and current physical activity in the elderly are associated with reductions in cardiovascular morbidity and mortality and improvement in or prevention of metabolic abnormalities that place elderly people at risk for these outcomes (CDC, 1989; van Dam et al., 2002; Lee et al., 2003; Esposito et al., 2003; Rosano et al., 2005). Based on these findings, the CDC currently recommends that elderly people engage in moderate-intensity physical activities such as bicycling on level terrain for 30 minutes or more at least five times a week in order to maintain their health (CDC, 1996).

While epidemiologic studies have produced compelling evidence for the health benefits provided by such moderate-intensity physical activities, it remains a largely open question to what extent more vigorous physical activities can offer additional benefits to the elderly. One of the main reasons for why this question has proven difficult to investigate lies in the lack of adequate statistical methods for estimating causal effects in this context. Current approaches in causal inference would define the causal effect of vigorous physical activity on a health outcome of interest by comparing the distribution of that outcome under the hypothetical scenario in which all subjects in the target population exercise at a given activity level to the corresponding distribution under the reference scenario in which all subjects abstain from vigorous physical activity. In order to estimate such treatment-specific counterfactual outcome distributions from observational data, however, one has to assume not only that the investigator has recorded all relevant confounding factors, but also that all subjects in the target population have a positive probability of selecting each of the treatment levels under consideration. Intuitively, this latter assumption of experimental treatment assignment (ETA) makes sense since we should not be able to estimate the counterfactual outcome distribution corresponding to a given treatment level if there exists a subgroup of the target population that in reality is never observed at that treatment level. In the context of studying the benefits of vigorous physical activity in the elderly, this assumption appears highly unrealistic since it can be expected that health problems would prevent a considerable proportion of subjects from participating in all but the lowest levels of vigorous physical activity. From a philosophical standpoint, it therefore does not even make sense to talk about the outcome distribution we would observe if all subjects were assigned to higher levels of vigorous physical activity. From a more practical standpoint, an analysis based on this approach would lead to an overestimate of the beneficial impact of higher levels of vigorous exercise since any estimate of the corresponding counterfactual distribution would be based solely on those subjects who are healthy enough to exercise at those levels.

van der Laan and Petersen (2007) recently proposed estimators of two kinds of causal effects that are defined on the basis of more realistic hypothetical scenarios. The first definition is based on realistic individualized treatment rules that, in contrast to the static rules described above, take into account a given subject's characteristics in order to assign a treatment level that is as close as possible to a specified target level while still remaining a realistic option for that subject. In the context of physical activity, for instance, we might consider hypothetical scenarios in which subjects are assigned to the highest vigorous activity level not exceeding a specified target level that they are still realistically capable of. The causal effect of vigorous physical activity could then be defined by comparing the outcome distribution we would observe for different target levels to the corresponding distribution we would observe under no vigorous physical activity. The second definition of causal effects is based on intention-to-treat rules that, like realistic individualized treatment rules, attempt to assign subjects to a specified target level, but allow subjects for whom this target level is not realistic to follow their self-selected treatment level rather than assigning them to the next highest realistic level. Causal effect estimates based on such rules thus aim to produce the results of an intention-to-treat analysis of a randomized trial in which a proportion of subjects fail to comply with treatment assignment and instead select their own treatment level. From a philosophical standpoint, causal effects defined on the basis of such realistic individualized treatment rules or intention-to-treat rules are appealing since the necessary counterfactual distributions are always well-defined. From a practical standpoint, analyses based on such rules offer the advantage of being protected from the bias that an analysis based on static treatment rules would be subject to if the ETA assumption is violated.

In this article, we present a data analysis examining the potential benefits of vigorous-intensity physical activity that compares the results obtained through a conventional analysis to those obtained by using the estimators developed in van der Laan and Petersen (2007). Our analysis illustrates that a conventional analysis based on static treatment rules yields severely biased results that dramatically

overestimate the true effect of higher levels of vigorous physical activity. At the same time, we show that causal effects based on realistic individualized treatment rules and intention-to-treat rules can be estimated without bias. The remainder of the article is organized as follows. After describing our data source, we briefly review the counterfactual framework for causal inference and describe the various estimators that have been proposed for estimating causal effects. We then present the details of our data analysis and close with a brief discussion of our results.

2 Data source

Tager et al. (1998) followed a group of people aged 55 years and older living in and around Sonoma, CA, over a time period of about ten years as part of a community-based longitudinal study of physical activity and fitness (Study of Physical Performance and Age Related Changes in Sonomans - SPPARCS). Our goal in analyzing the data that were collected as part of this study is to examine the effect of vigorous LTPA as recorded at the baseline interview on subsequent five-year all-cause mortality.

Our measure of vigorous LTPA is defined based on a questionnaire in which participants were asked how many hours during the past seven days they had participated in twelve common vigorous physical activities such as jogging, swimming, bicycling on hills, or racquetball. Activities were assigned standard intensity values in metabolic equivalents (METs) (Ainsworth et al., 1993); one MET approximately equals the oxygen consumption required for sitting quietly. A continuous summary score was obtained by multiplying these intensity values by the number of hours engaged in the various activities and summing up over all activities considered here. The treatment variable A was then defined as a categorical version of this summary LTPA score:

$$A = \begin{cases} 0 & \text{if } LTPA = 0 \text{ METs} \\ 1 & \text{if } 0 \text{ METs} < LTPA \leq 10 \text{ METs} \\ 2 & \text{if } 10 \text{ METs} < LTPA \leq 20 \text{ METs} \\ 3 & \text{if } 20 \text{ METs} < LTPA \leq 40 \text{ METs} \\ 4 & \text{if } 40 \text{ METs} < LTPA \leq 60 \text{ METs} \\ 5 & \text{if } 60 \text{ METs} < LTPA \end{cases} \tag{1}$$

To compare, the current CDC recommendation for engaging in moderate-intensity physical activity for 30 minutes at least five times a week corresponds to an energy expenditure of 22.5 METs.

Apart from sex and age, the primary confounding factor of the relationship between LTPA and all-cause mortality is likely to be given by a subject's underlying level of general health. Healthier subjects will not only tend to experience lower mortality risks, but are also more likely to engage in higher levels of vigorous physical activity. To control for this source of confounding, our analysis adjusts for a number of covariates that are intended to capture a subject's underlying level of health. Participants were asked, for instance, to rate their health as excellent, good, fair, or poor. Self-reported physical functioning was defined from a series of questions, originally developed by Nagi (1976) and Rosow and Breslau (1966), that assessed the degree of difficulty a participant experienced in various activities of daily living. On the basis of this questionnaire, we classified a participant's level of physical functioning as excellent, moderately impaired, or severely impaired. In addition, participants were asked about the previous occurrence of cardiac events such as myocardial infarctions, the presence of a number of chronic health conditions, their smoking status, as well as a possible decline in physical activity compared to 5 or 10 years earlier. Table 1 summarizes the definition of the covariates we adjust for as potential confounding factors.

Of the 2092 participants enrolled in the SPPARCS study, 15 did not answer all the questions needed to define their level of vigorous physical activity; an additional 26 were missing information about at a least one of the confounding factors described above. Our analysis is based on the remaining 2051 participants. We note that the outcome of interest, five-year survival status, was available for all study participants so that we do not have to adjust for right censoring.

Table 1: Definition of indicator variables that are considered as potential confounders.

Variable	Definition
$FEMALE$	Female
$AGE.1$	$\leq$ 60 years old
$AGE.2$	60-70 years old
$AGE.4$	80-90 years old
$AGE.5$	90-100 years old
$HTL.EX$	Excellent self-rated health
$HLT.FAIR$	Fair self-rated health
$HLT.POOR$	Poor self-rated health
$NRB.FAIR$	Moderately impaired physical functioning ($0.5 \leq$ NRB score ¡ 1.0)
$NRB.POOR$	Severely impaired physical functioning (NRB score ¡ 0.5)
$CARD$	Previous occurrence of any of the following cardiac events: Angina, myocardial infarction, congestive heart failure, coronary by-pass surgery, and coronary angioplasty
$CHRON$	Presence of any of the following chronic health conditions: stroke, cancer, liver disease, kidney disease, Parkinson's disease, and diabetes mellitus
$SMK.CURR$	Current smoker
$SMK.EX$	Former smoker
$DECLINE$	Activity decline compared to 5 or 10 years earlier

3 Methods

The observed data are given by n i.i.d. copies of $O = (W, A, Y)$, where W denotes the collection of adjustment variables, A gives the categorical physical activity level, and Y is an indicator for death in the five years following the baseline interview. Within the counterfactual framework for causal inference, as first introduced by Neyman (1923) and further developed by Rubin (1978) and Robins (1986, 1987), this observed data structure O is viewed as a censored version of a hypothetical full data structure $X = (Y_a : a \in \mathcal{A})$ that contains the outcome Y_a we would have observed on this subject had she been assigned to treatment level a for all a in the collection $\mathcal{A} = \{0, 1, \ldots, 5\}$ of possible treatment levels. The causal effect of vigorous physical activity on all-cause mortality could now be defined by comparing the mortality risk $E[Y_a]$ we would observe if all subjects in the target population exercised at a given level $a > 0$ to the corresponding mortality risk $E[Y_0]$ we would observe if all subjects abstained from vigorous physical activity.

As mentioned previously, such mean counterfactual outcomes can only be estimated from the observed data if the investigator has recorded all relevant confounding factors and if all subjects in the target population have positive probability of selecting each of the treatment levels. This latter assumption of experimental treatment assignment can be formalized by requiring that for all candidate static treatment interventions $a = 0, 1, \ldots, 5$, we have with probability 1.0 that

$$g(a \mid W) \equiv P(A = a \mid W) > 0. \tag{2}$$

In fact, it has been shown that estimation of mean counterfactual outcomes becomes problematic even if there exist values of a and W for which the treatment assignment probabilities $g(a \mid W)$ are not identically equal to zero, but very close to zero (Neugebauer and van der Laan, 2005). To avoid problems due to such a practical violation of the ETA assumption, we may hence require in practice that, for $a = 0, 1, \ldots, 5$, we have $g(a \mid W) > \alpha$ with probability 1.0, with $\alpha = 0.05$, for instance.

Estimators of causal effects defined on the basis of the realistic individualized treatment rules discussed in van der Laan and Petersen (2007) do not rely on the ETA assumption. Given a target treatment level a and a subject's baseline covariates W, such rules assign the highest treatment level not exceeding a that the subject is still realistically capable of. Specifically, let

$$\mathcal{D}(W) = \{a \in \mathcal{A} : g(a \mid W) \geq \alpha\} \tag{3}$$

denote the set of treatment options that, given baseline covariates W, are realistic for a particular subject in the sense that she would select any one of those treatment options with a probability of at least α. A realistic individualized treatment rule can then be defined as

$$d(a, W) = \max\{a^* \in \mathcal{D}(W) : a^* \leq a\}. \tag{4}$$

As with static treatment regimens, we use the notation $Y_{d(a,W)}$ to denote the outcome we would have observed on the subject had she followed the individualized rule $d(a, W)$, i.e. $Y_{d(a,W)} \equiv Y_{\tilde{a}}$ where $\tilde{a} = d(a, W)$. A realistic causal effect of vigorous physical activity on all-cause mortality can now be defined by comparing the mortality risk $E[Y_{d(a,W)}]$ we would observe if all subjects in the target population followed a given rule $d(a, W)$, $a > 0$, to the corresponding mortality risk $E[Y_{d(0,W)}] = E[Y_0]$ we would observe if all subjects abstained from vigorous physical activity. By the definition of $d(a, W)$, we have, for $a = 0, 1, \ldots, 5$, that $g(d(a, W) \mid W) > \alpha$ with probability 1.0, demonstrating that the equivalent of assumption (2) is trivially satisfied in estimating the corresponding causal effects.

Under an intention-to-treat rule $d(a, A, W)$, subjects are assigned to a specified target treatment level a if that treatment level represents a realistic option for them, but are allowed to follow their self-selected treatment A otherwise:

$$d(a, A, W) = I(a \in \mathcal{D}(W))a + I(a \notin \mathcal{D}(W))A. \tag{5}$$

An intention-to-treat causal effect of vigorous physical activity on all-cause mortality can now be defined by comparing the counterfactual mortality risks $E[Y_{d(a,A,W)}]$, $a > 0$, and $E[Y_{d(0,A,W)}] = E[Y_0]$. Note that we have

$$E[Y_{d(a,A,W)}] = E\Big[Y_a I(a \in \mathcal{D}(W))\Big] + E\Big[Y I(a \notin \mathcal{D}(W))\Big]. \tag{6}$$

The second quantity is trivially identified by the observed data, and $a \in \mathcal{D}(W)$ guarantees that $g(a \mid W) > \alpha$ with probability 1.0, ensuring identifiability of the second quantity, so that the equivalent of assumption (2) is guaranteed to hold in the estimation of intention-to-treat causal effects. We note that the true treatment mechanism g and therefore also the set $\mathcal{D}(W)$ of realistic treatment options will generally be unknown. In practice, it will therefore usually be necessary to substitute a given estimate g^* of the treatment mechanism g in the definition of $D(W)$.

Four different classes of estimators have been proposed for estimating mean counterfactual outcomes corresponding to static treatment rules: G-computation estimators (Robins, 1986), inverse-probability-of-treatment-weighted (IPTW) estimators (Robins, 2000), double robust IPTW (DR-IPTW) estimators (van der Laan and Robins, 2003), and targeted maximum-likelihood estimators (van der Laan and Rubin, 2006), with natural analogues of all of these estimators in the context of realistic individualized treatment rules and intention-to-treat rules. While it is well-known that the IPTW estimator can suffer from considerable bias if the ETA assumption is violated, the remaining three estimators are in fact also severely compromised in such situations in that they now have to rely fully on model assumptions that cannot be tested from the data (Neugebauer and van der Laan, 2005). Since this latter phenomenon is rarely discussed in the literature, we will provide a practical illustration by comparing the estimates obtained by each of these four estimators for the three different causal effects defined above. We next review the definition and implementation of these four estimators in order to be able to discuss their behavior in more detail.

We begin with estimators of the mean counterfactual outcome $\psi = E[Y_{d(a,W)}]$ for a given realistic individualized treatment rule $d(a, W)$. Note that the mean counterfactual outcome $E[Y_a]$ for a given static treatment rule corresponds to the special case of setting $\alpha = 0$ in the definition of $\mathcal{D}(W)$. The G-computation estimator of ψ is based on the observation that under the assumption of no unmeasured confounders, this parameter is identified by the observed data as

$$\psi = E[Y_{d(a,W)}] = E_W\Big[E[Y \mid A = a, W]\Big]. \tag{7}$$

This immediately implies a substitution estimator based on estimates of the marginal distribution of W, $P(W)$, and the conditional distribution of Y given A and W, $P(Y \mid A, W)$. The first distribution can be estimated non-parametrically by the empirical distribution of W in our sample, but estimation of $P(Y \mid A, W)$ will generally require specification of a parametric model. In the case of a binary outcome Y, an estimate Q_n of the regression $Q(A, W) = E[Y \mid A, W]$ based on an appropriate logistic regression model completely defines an estimate of the conditional distribution $P(Y \mid A, W)$. The corresponding

substitution estimator for ψ is then given by

$$\psi_n^{G-comp} = \frac{1}{n}\sum_{i=1}^{n} Q_n(d(a, W_i), W_i). \tag{8}$$

This estimator gives a consistent estimate of ψ if the model for $Q(A, W)$ is correctly specified.

The IPTW and DR-IPTW estimators are based on the general estimating function methodology described in van der Laan and Robins (2003) that is based on the following three steps. First, estimating functions for ψ are obtained assuming that we have access to the full data structure X. These estimating functions are then mapped into functions of the observed data structure by applying an IPTW mapping. Lastly, a class of more robust and efficient estimating functions is obtained by subtracting from these IPTW estimating functions their projection onto the tangent space for the treatment mechanism in the model that only makes the assumption of no unmeasured confounders. In a non-parametric model, the only unbiased full-data estimating function for ψ is given by

$$D^{Full}(X \mid \psi) = Y_{d(a,W)} - \psi. \tag{9}$$

A corresponding IPTW estimating function is given by

$$D^{IPTW}(O \mid g, \psi) = \frac{I(A = d(a, W))}{g(A \mid W)} Y - \psi. \tag{10}$$

The IPTW estimator ψ_n^{IPTW} is defined as the solution of the estimating equation

$$0 = \frac{1}{n}\sum_{i=1}^{n} D^{IPTW}(O_i \mid g_n, \psi), \tag{11}$$

where g_n is an estimate of g that may, for example, be obtained as the maximum-likelihood estimate of g in an appropriately specified parametric model. Specifically, this estimator is given by

$$\psi_n^{IPTW} = \frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i = d(a, W_i))}{g_n(A_i \mid W_i)} Y_i. \tag{12}$$

It gives a consistent estimate of ψ if the model for the treatment mechanism g is correctly specified.

The projection of D^{IPTW} onto the nuisance tangent space T_{NUC} corresponding to the treatment mechanism under the assumption of no unmeasured confounders can be computed as

$$\begin{aligned} \Pi[D^{IPTW} \mid T_{NUC}] &= E[D^{IPTW} \mid A, W] - E[D^{IPTW} \mid W] \\ &= \frac{I(A = d(a, W))}{g(A \mid W)} Q(A, W) - Q(d(a, W), W) \end{aligned}$$

so that the DR-IPTW estimating function is given by

$$D^{DR}(O \mid g, Q, \psi) = \frac{I(A = d(a, W))}{g(A \mid W)} \Big[Y - Q(A, W)\Big] + Q(d(a, W), W) - \psi. \tag{13}$$

The corresponding DR-IPTW estimator ψ_n^{DR} is defined as the solution of the estimating equation

$$0 = \frac{1}{n}\sum_{i=1}^{n} D^{DR}(O_i \mid g_n, Q_n, \psi). \tag{14}$$

Specifically,

$$\psi_n^{DR} = \frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i = d(a, W_i))}{g_n(A_i \mid W_i)} \Big[Y_i - Q_n(A_i, W_i)\Big] + Q_n(d(a, W_i), W_i). \tag{15}$$

This estimator gives a consistent estimate of ψ if the model for either g or Q is correctly specified. It is also locally efficient in the sense that correct specification of both models yields an efficient estimator.

Like the G-computation estimator, the targeted MLE of ψ is a substitution estimator based on estimates of the components $P(W)$ and $P(Y \mid A, W)$ of the observed data density. In order to avoid relying on an *a priori* specified parametric model for the latter component, we may often want to

employ a data-adaptive model selection approach such as the Deletion/Substituion/Addition algorithm (Sinisi and van der Laan, 2004) or Least Angle Regression (Efron et al., 2004) for the purposes of estimating this conditional density. This is somewhat problematic, however, since such algorithms will select an appropriate model based on a criterion that is aimed at estimating the nuisance parameter $P(Y \mid A, W)$ efficiently, which in general does not lead to an efficient estimator of the parameter of interest ψ. The targeted MLE therefore first updates the initial estimate of the observed-data density that would be used by the G-computation estimator in a way that targets estimation of this density at the parameter of interest and makes the corresponding substitution estimator double robust and locally efficient. Specifically, this is achieved by formulating a parametric model indexed by a Euclidean parameter ϵ through the initial estimate of the observed-data density at $\epsilon = 0$ whose scores include the components of the efficient influence curve of ψ at the initial density estimate, obtaining a maximum-likelihood estimate of ϵ in this model, and updating the original density estimate accordingly.

Since this targeted maximum-likelihood approach was only recently developed by van der Laan and Rubin (2006), we will illustrate it here in the context of estimating the parameter of interest ψ. For this purpose, let P_n^0 be an initial estimator of the observed-data density that estimates the marginal distribution of W by the empirical distribution of W, the treatment mechanism g by an estimate $g(P_n^0)$, and the conditional distribution of Y given A and W by an initial fit $Q(P_n^0)$ that can be represented in the form of the logistic function

$$Q(P_n^0)(A, W) = \frac{1}{1 + \exp(-m_n^0(A, W))}. \tag{16}$$

We then need to formulate a parametric fluctuation through this initial density estimate whose scores at the initial estimate include the components of the efficient influence curve for ψ. This efficient influence curve, given by the influence curve $D(P)$ of the DR-IPTW estimator

$$D(P) = \frac{I(A = d(a, W))}{g(A \mid W)}\Big[Y - Q(A, W)\Big] + Q(d(a, W), W) - \psi, \tag{17}$$

can be decomposed as

$$\begin{aligned} D(P) \quad = \quad & D(P) - E[D(P) \mid A, W] + \\ & E[D(P) \mid A, W] - E[D(P) \mid W] + \\ & E[D(P) \mid W] - E[D(P)], \end{aligned} \tag{18}$$

corresponding to scores for $P(Y \mid A, W)$, $P(A \mid W)$, and $P(W)$, respectively. Specifically, we have that

$$\begin{aligned} D_1(P) &= D(P) - E[D(P) \mid A, W] \\ &= \frac{I(A = d(a, W))}{g(A \mid W)}\Big[Y - Q(A, W)\Big] \qquad (19) \\ D_2(P) &= E[D(P) \mid A, W] - E[D(P) \mid W] \\ &= 0 \qquad (20) \\ D_3(P) &= E[D(P) \mid W] - E[D(P)] \\ &= Q(d(a, W), W) - \psi_a. \qquad (21) \end{aligned}$$

Since the empirical distribution of W is a non-parametric maximum-likelihood estimator of $P(W)$, it in particular equals the MLE of $P(W)$ in any parametric fluctuation through this initial estimate so that we do not need to concern ourselves with updating this component of the observed-data density. Since the parameter of interest is orthogonal to the treatment mechanism g so that $D_2(P) = 0$, we also do not need to obtain an update of an initial estimate of g. As a submodel through $P_n^0(Y \mid A, W)$, we will consider a logistic regression model that is identical to the initial fit $Q(P_n^0)$ except for an added covariate $h(P_n^0)(A, W)$:

$$Q(P_n^0)(\epsilon)(A, W) = \frac{1}{1 + \exp(-m_n^0(A, W) - \epsilon h(P_n^0)(A, W))} \tag{22}$$

The covariate $h(P_n^0)(A, W)$ needs to be chosen such that the score of this submodel at $\epsilon = 0$ is equal to $D_1(P_n^0)$, the component of the efficient influence curve corresponding to $P(Y \mid A, W)$ at the initial density estimate. The score of the selected submodel at $\epsilon = 0$ is given by

$$S(0) = h(P_n^0)(A, W)\Big(Y - Q(P_n^0)(A, W)\Big). \tag{23}$$

Solving for h such that

$$\begin{aligned} S(0) &= D_1(P_n^0) \\ &= \frac{I(A = d(a,W))}{g(P_n^0)(A \mid W)}\Big[Y - Q(P_n^0)(A,W)\Big] \end{aligned} \tag{24}$$

yields the solution

$$h(P_n^0)(A,W) = \frac{I(A = d(a,W))}{g(P_n^0)(A \mid W)}. \tag{25}$$

Let ϵ_n denote the MLE of ϵ in $Q(P_n^0)(\epsilon)$, which can be obtained by simply regressing Y on $h(P_n^0)(A,W)$ according to a logistic regression model with offset equal to $m_n^0(A,W)$. The targeted MLE of ψ is then given by the substitution estimator based on the updated estimate

$$Q_n^1(A,W) = \frac{1}{1 + \exp(-m_n^0(A,W) - \epsilon_n h(P_n^0)(A,W))} \tag{26}$$

of the regression $Q(A,W)$. Specifically, we have that

$$\psi_n^{tMLE} = \frac{1}{n}\sum_{i=1}^{n} Q_n^1(d(a,W_i),W_i). \tag{27}$$

To summarize, implementing this estimator thus requires initial estimates of the regression Q and the treatment mechanism g as they would also be used by the three estimators described above, updating the estimate for Q in a simple univariate logistic regression, and then computing the corresponding substitution estimator of ψ. The resulting targeted MLE solves the double robust estimating equation based on $Q_n^1(A,W)$ and g_n, i.e.

$$\frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i = d(a,W_i))}{g(P_n^0)(A_i \mid W_i)}\Big[Y_i - Q_n^1(A_i,W_i)\Big] + Q_n^1(d(a,W_i),W_i) - \psi_n^{tMLE} = 0, \tag{28}$$

so that it is in fact equivalent to the DR-IPTW estimator given in (15) with $Q_n^1(A,W)$ substituted for $Q_n(A,W)$. Like the DR-IPTW estimator, the targeted MLE is therefore consistent if at least one of the two nuisance parameters g and Q is estimated consistently. Similarly, the estimator is locally efficient in the sense that it is efficient if both of these nuisance parameters are estimated consistently.

As mentioned previously, estimation of the mean counterfactual outcome $E[Y_a]$ corresponding to a static treatment intervention represents a special case of the realistic individualized treatment rules considered here. G-computation, IPTW, and DR-IPTW estimators of the mean counterfactual outcome $\phi \equiv E[Y_{d(a,A,W)}]$ corresponding to an intention-to-treat rule are straightforward to derive and are presented in van der Laan and Petersen (2007). In order to obtain a targeted MLE of ϕ, we can use that by (6) the efficient influence curve of ϕ in a non-parametric model can be written as the sum of the efficient influence curve of a non-parametric estimator of $\phi_1 = E[YI(a \notin \mathcal{D})]$ and the efficient influence curve of a non-parametric estimator of $\phi_2 = E[Y_a I(a \in \mathcal{D})]$. These are given by

$$D^1(P) = I(a \notin \mathcal{D})Y - \phi_1 \tag{29}$$

and

$$D^2(P) = I(a \in \mathcal{D})\left\{\frac{I(A=a)}{g(A \mid W)}\Big[Y - Q(A,W)\Big] + Q(a,W)\right\} - \phi_2, \tag{30}$$

respectively, yielding

$$D(P) = I(a \notin \mathcal{D})Y + I(a \in \mathcal{D})\left\{\frac{I(A=a)}{g(A \mid W)}\Big[Y - Q(A,W)\Big] + Q(a,W)\right\} - \phi \tag{31}$$

as the efficient influence curve for ϕ. The component of this influence curve corresponding to the score for $P(Y \mid A,W)$ is given by

$$\begin{aligned} D(P) - E[D(P) \mid A,W] &= I(a \notin \mathcal{D})\Big[Y - Q(A,W)\Big] + I(a \in \mathcal{D})\left\{\frac{I(A=a)}{g(A \mid W)}\Big[Y - Q(A,W)\Big]\right\} \\ &= \left\{I(a \notin \mathcal{D}) + I(a \in \mathcal{D})\frac{I(A=a)}{g(A \mid W)}\right\}\Big[Y - Q(A,W)\Big]. \end{aligned} \tag{32}$$

The covariate $h(P_n^0)(A,W)$ needed for the univariate regression to update the initial fit for Q is thus given by

$$h(P_n^0)(A,W) = I(a \notin \mathcal{D}) + I(a \in \mathcal{D})\frac{I(A=a)}{g(P_n^0)(A \mid W)}. \quad (33)$$

The problems arising if the ETA assumption is violated are most clearly seen in the case of the IPTW estimator. By downweighting observations that were likely to have received their observed treatment and upweighting those that were instead unlikely to have received their observed treatment, this estimator essentially works by creating a new sample in which treatment assignment is independent of the baseline covariates. This approach breaks down if a subgroup of the target population never selects some of candidate treatment levels. If older, less healthy subjects, for example, are never observed to participate in high levels of vigorous physical activity, none of the subjects in the corresponding re-weighted sample will be older and less healthy, leading to an underestimate of the corresponding counterfactual mortality risk under high levels of vigorous physical activity.

In the same situation, the G-computation estimator has to rely entirely on model assumptions that cannot be tested from the observed data. Since older, less healthy subjects are never observed at higher levels of vigorous physical activity, their conditional mean outcome $E[Y \mid A,W]$ for these exercise levels is undefined. A corresponding estimate can never be obtained from the observed data unless one is willing to extrapolate from the conditional mean outcomes estimated for other values of A and W. To illustrate this point, consider the simplified example in which A is a binary indicator for a high level of vigorous physical activity and W is an indicator for poor health. Then none of the subjects in our target population might fall in the group with $W = 1$ and $A = 1$ so that $E[Y \mid A = 1, W = 1]$ is undefined. In order to still obtain an estimate of this quantity, we would be forced to assume an additive model for Q according to which $Q(A,W) = \beta_0 + \beta_1 A + \beta_2 W$. Since the non-parametric model $Q(A,W) = \beta_0 + \beta_1 A + \beta_2 W + \beta_3 A \times W$ is not identifiable, this assumption of no interaction between A and W cannot be tested from the observed data.

Like the G-computation estimator, the DR-IPTW estimator and the targeted MLE rely entirely on extrapolation through Q if the ETA assumption is violated. To complicate matters, however, they also require that the estimate of g is based on a model for the treatment mechanism that satisfies the ETA assumption, i.e. the model for g must in fact be mis-specified (van der Laan and Robins, 2003). In summary, all four estimators of causal effects are thus severely compromised if the ETA assumption does not hold, illustrating that the solution in such cases does not lie in turning to the G-computation or DR-IPTW estimators for which the resulting problems are not as immediately apparent as for the IPTW estimator, but in focusing on realistically defined causal effects that are guaranteed to be identified from the observed data.

4 Results

The treatment mechanism was estimated by a multinomial regression model that included main-effect terms for all indicator variables defined in table 1. The regression $E[Y \mid A,W]$ was similarly estimated by a logistic regression model that included these same main-effect terms as well as indicator variables for the treatment categories 1 through 5. We evaluated the goodness-of-fit of this latter model using the Hosmer-Le Cessie test introduced by Hosmer et al. (1997) as an improvement of the Hosmer-Lemeshow test (Hosmer and Lemeshow, 1980). This test yielded a p-value of 0.10, providing little evidence against the assumption that this model adequately describes the data. To evaluate the fit of our treatment model, we followed the advise of Hosmer and Lemeshow (2000) and treated this model as a set of independent binary logistic regression models of each treatment category against the remaining categories. Applying the Hosmer-Le Cessie test to each of these binary logistic regression models, we obtained p-values of 0.51, 0.54, 0.33, 0.27, 0.78, and 0.94, suggesting that the treatment model fits the data quite well.

Tables 2 and 3 summarize the fits we obtained for g and Q, respectively. The treatment fit reveals a clear violation of the ETA assumption: No subjects in the oldest age group (90-100 years) are observed at the treatment levels $A = 3$ and $A = 5$. Likewise, no subjects with poor self-rated health are observed at the treatment levels $A = 4$ and $A = 5$. In addition, subjects with severely impaired physical functioning are very unlikely to follow treatments $A = 4$ and $A = 5$. The fit we obtained for Q indicates that these three groups of subjects are at considerably increased risks of mortality, suggesting that estimates of the counterfactual mortality risks for the higher three treatment categories will be biased low. Since the

DR-IPTW estimator and the targeted MLE both require an estimate of the treatment mechanism that satisfies the ETA assumption, fitted treatment assignment probabilities below 0.05 were set to 0.05.

Table 2: Treatment model fit. The entries in the first column give the factor by which the relative risk of falling in category A=1 rather than A=0 changes when the covariate under consideration is changed from 0 to 1. Entries in the remaining columns are interpreted accordingly.

	A=1	A=2	A=3	A=4	A=5
AGE.1	1.16	1.57	1.37	1.32	1.44
AGE.2	1.37	1.57	1.47	1.32	1.37
AGE.4	0.74	0.94	0.83	0.83	1.02
AGE.5	0.24	1.03	0.00	1.04	0.00
HLT.EX	1.09	1.10	1.46	1.29	1.67
HLT.FAIR	0.56	0.58	0.47	0.39	0.45
HLT.POOR	0.50	0.43	0.33	0.00	0.00
NRB.POOR	0.55	0.40	0.29	0.07	0.17
NRB.FAIR	0.78	0.82	0.70	0.99	0.53
SMOKE.CURR	0.65	0.43	0.32	0.61	0.33
SMOKE.EX	1.00	1.23	1.09	1.25	1.20
CARD	0.90	1.29	1.18	0.89	1.46
CHRONIC	1.19	1.14	1.13	1.11	0.93
FEMALE	0.94	0.86	0.82	0.89	0.55
DECLINE	0.67	0.39	0.52	0.37	0.33

Table 3: Fit for Q. Estimated odds ratios for mortality along with 95% confidence intervals and p-values.

	OR	95% CI	p-value
AGE.1	0.12	(0.05, 0.31)	¡10e-4
AGE.2	0.43	(0.29, 0.64)	¡10e-4
AGE.4	3.41	(2.34, 4.96)	¡10e-4
AGE.5	5.74	(2.07, 15.91)	¡10e-4
HLT.EX	0.76	(0.50, 1.16)	0.2039
HLT.FAIR	2.01	(1.39, 2.93)	¡10e-4
HLT.POOR	2.84	(1.51, 5.34)	0.0012
NRB.POOR	1.94	(1.21, 3.13)	0.0063
NRB.FAIR	0.89	(0.61, 1.29)	0.5279
SMOKE.CURR	3.73	(2.22, 6.29)	¡10e-4
SMOKE.EX	1.38	(0.99, 1.94)	0.0584
CARD	1.60	(1.13, 2.26)	0.0080
CHRONIC	1.44	(1.06, 1.95)	0.0204
FEMALE	0.52	(0.37, 0.72)	¡10e-4
DECLINE	1.46	(1.05, 2.05)	0.0266
A=1	0.86	(0.55, 1.34)	0.5072
A=2	0.81	(0.51, 1.29)	0.3849
A=3	0.78	(0.47, 1.29)	0.3360
A=4	0.45	(0.18, 1.09)	0.0770
A=5	0.80	(0.37, 1.76)	0.5866

Tables 4 and 5 summarize the realistic indvidualized treatment rule and the intention-to-treat rule. Both tables show that only about 50% of all subjects are estimated to be capable of engaging in the highest level of vigorous physical activity. Likewise, only about 75% of all subjects are estimated to be capable of the second highest level. These observations further underscore the severity of the ETA violation encountered in this data set. In comparing tables 4 and 5, we note that the intention-to-treat causal effects of high levels of vigorous physical activity are likely to be smaller than the corresponding

realistic causal effects. Under the intention-to-treat rule $d(5, A, W)$, close to 25% of all subjects are assigned to the lowest treatment level $A = 0$ while the corresponding realistic individualized treatment rule $d(5, W)$ assigns no subjects to $A = 0$. In general, the realistic individualized treatment rule results in treatment assignments closer to the specified target level than those obtained from the intention-to-treat rule. In addition, the latter rule produces a few cases in which subjects are assigned to treatment levels that exceed the given target level. For the sake of estimating the causal effect of vigorous physical activity, these observations would seem to make the realistic individualized treatment rule a somewhat more appealing option than the intention-to-treat rule.

Table 4: The realistic individualized treatment rule. A given row shows the treatment levels $\tilde{a} \equiv d(a, W)$ that subjects were actually assigned to when the target level was set at a.

	$\tilde{a} = 0$	$\tilde{a} = 1$	$\tilde{a} = 2$	$\tilde{a} = 3$	$\tilde{a} = 4$	$\tilde{a} = 5$
$a = 0$	2051	0	0	0	0	0
$a = 1$	11	2040	0	0	0	0
$a = 2$	0	41	2010	0	0	0
$a = 3$	0	41	97	1913	0	0
$a = 4$	0	41	91	441	1478	0
$a = 5$	0	41	91	381	454	1084

Table 5: The intention-to-treat treatment rule. A given row shows the treatment levels $\tilde{a} \equiv d(a, A, W)$ that subjects were actually assigned to when the target level was set at a.

	$\tilde{a} = 0$	$\tilde{a} = 1$	$\tilde{a} = 2$	$\tilde{a} = 3$	$\tilde{a} = 4$	$\tilde{a} = 5$
$a = 0$	2051	0	0	0	0	0
$a = 1$	11	2040	0	0	0	0
$a = 2$	35	3	2011	1	1	0
$a = 3$	108	16	7	1918	2	0
$a = 4$	338	88	66	56	1491	12
$a = 5$	492	161	134	110	45	1109

As argued above, the lack of non-parametric identifiability of causal parameters under a violation of the ETA assumption is most easily seen in the case of the IPTW estimator which is likely to suffer from considerable bias. Wang et al. (2006) propose the following simulation-based approach for obtaining an estimate of this bias: Given estimates of $P(W)$, g, and Q, we can simulate realizations of the observed data structure. For this estimated data-generating distribution, the true parameter values for the parameters of interest can be computed through G-computation. At the same time, we can obtain a sampling distribution of IPTW estimates by applying the IPTW estimator to a large number of simulated realizations of the observed data structure. Since the assumption of no unmeasured confounders is trivially satisfied in this simulation study, any discrepancy between the mean of these estimates and the true parameter value must reflect a violation of the ETA assumption. Table 6 summarizes the estimated bias of the IPTW estimator of the counterfactual mortality risk for each of the three different kinds of causal effects. The table shows that the IPTW estimator dramatically underestimates the counterfactual mortality risk for static treatment interventions at the highest two activity levels, with considerable problems even for the third highest level of activity. These observations are in agreement with our earlier arguments according which a lack of older and less healthy subjects among the higher activity levels should lead to an underestimate of the corresponding mortality risks. In contrast, table 6 shows only a negligible bias for estimating such risks on the basis of realistic individualized treatment rules and intention-to-treat rules. We stress that this diagnostic simulation should be interpreted to give not only an estimate of the bias seen in the IPTW estimator, but, more generally, a sense of the extent to which an ETA violation makes the causal parameters of interest non-parametrically non-identifiable. In the present case, for instance, we would therefore also want to treat any estimates of static causal effects offered by the G-computation, DR-IPTW, and targeted maximum-likelihood estimators as unreliable and potentially misleading.

Given the counterfactual mortality risk estimators described in section 3, estimators of the relative

Table 6: Estimated ETA bias for the IPTW estimator of the counterfactual mortality risk as a percentage of the true parameter value.

	Static	Realistic	ITT
A=0	-0.23%	-0.23%	-0.23%
A=1	-2.63%	0.05%	-0.03%
A=2	-4.94%	0.04%	0.13%
A=3	-14.45%	0.22%	0.20%
A=4	-48.75%	1.16%	1.05%
A=5	-50.54%	-0.18%	0.11%

risk (relative to $A = 0$) are straightforward to obtain for the G-computation, IPTW, and DR-IPTW estimators by simply dividing the corresponding two mortality risk estimators. Since the targeted MLE is always aimed at a particular parameter of interest, this simple approach does not work for obtaining a targeted MLE of the relative risk of mortality. Section A in the appendix shows that this task is still fairly straightforward, however, given the work we have already done in section 3. Table 7 summarizes the relative risk estimates for the three different kinds of causal effects obtained by the four different estimators.

In the analysis based on static treatment interventions, the IPTW estimator appears to provide strong evidence for a protective effect of vigorous physical activity at the highest two levels, with an estimated 4-fold reduction in risk for the second-highest level. The realistic and intention-to-treat analysis, however, provide much weaker evidence for such a protective effect. As expected, the intention-to-treat causal effect estimates tend to be closer to the null value than the corresponding realistic estimates. Given the results of the simulation study summarized in table 6, we are led to conclude that the IPTW estimates based on static treatment interventions dramatically overstate the beneficial impact of high levels of vigorous physical activity.

The remaining three estimators likewise tend to estimate stronger reductions in risk in the static analysis than in the realistic and intention-to-treat analyses, with both the DR-IPTW estimator and the targeted MLE indicating a significant protective effect for $A = 4$ in the static analysis that becomes non-significant in the realistic and intention-to-treat analyses. Interestingly, the G-computation estimator also yields a smaller estimated reduction in risk for $A = 4$ in the latter two analyses than in the former one, but tighter confidence intervals for the realistic and intention-to-treat analyses actually make the corresponding causal effect estimates significant while this is not the case in the static analysis. We speculate that the greater sampling variability observed in the static analysis is likely a result of the extrapolation that is required to estimate the expected mortality outcome for a large number of subjects that are never observed at the highest two treatment levels. For all four estimators, the static analysis suggest a markedly greater mortality risk for $A = 5$ than for $A = 4$, a finding that would be quite hard to interpret. The remaining two analyses, in contrast, provide much more compatible estimates for these two activity levels. These observations lend credence to the idea that the static effect estimates not only of the IPTW estimator, but also of the G-computation, DR-IPTW, and targeted maximum-likelihood estimator ought to be treated as unreliable and potentially misleading. On the basis of the more trustworthy realistic and intention-to-treat analyses, the data suggest that high levels of vigorous physical activity may confer reductions in mortality risk on the order of 15-30%, although in most cases the evidence for such an effect does not quite reach the 0.05 level of significance.

5 Discussion

The data analysis presented in this article illustrates the problems encountered in attempting to estimate the causal effect of a static treatment intervention if the ETA assumption is violated. While it is fairly well-known that such a violation can cause strong bias in the IPTW estimator, its effects on other estimators of static causal effects have received little attention in the literature. With the G-computation estimator, the DR-IPTW estimator, and the targeted MLE all relying on extrapolation from a correctly specified model for Q and the latter two estimators in addition requiring a mis-specified model for the

Table 7: Estimates of the relative risk of mortality (relative to $A = 0$) along with 95% confidence intervals based on the bootstrap.

	G-comp	IPTW	DR-IPTW	tMLE
Static				
A=1	0.90 (0.65, 1.20)	0.97 (0.68, 1.29)	0.96 (0.69, 1.28)	0.96 (0.69, 1.28)
A=2	0.91 (0.64, 1.23)	0.90 (0.60, 1.22)	0.92 (0.63, 1.27)	0.93 (0.63, 1.30)
A=3	0.88 (0.59, 1.21)	0.77 (0.44, 1.07)	0.84 (0.56, 1.14)	0.87 (0.58, 1.18)
A=4	0.59 (0.22, 1.01)	0.23 (0.06, 0.43)	0.52 (0.20, 0.92)	0.48 (0.15, 0.88)
A=5	0.86 (0.43, 1.35)	0.55 (0.21, 0.90)	0.97 (0.48, 1.50)	1.05 (0.53, 1.60)
Realistic				
A=1	0.91 (0.66, 1.19)	1.00 (0.72, 1.32)	0.95 (0.70, 1.28)	0.95 (0.70, 1.28)
A=2	0.87 (0.63, 1.17)	0.97 (0.67, 1.34)	0.99 (0.66, 1.30)	1.00 (0.66, 1.32)
A=3	0.85 (0.62, 1.13)	0.81 (0.50, 1.22)	0.91 (0.59, 1.22)	0.91 (0.58, 1.23)
A=4	0.73 (0.53, 0.97)	0.58 (0.34, 1.06)	0.69 (0.40, 1.05)	0.69 (0.41, 1.05)
A=5	0.81 (0.60, 1.06)	0.66 (0.38, 1.19)	0.78 (0.47, 1.17)	0.78 (0.46, 1.20)
ITT				
A=1	0.91 (0.66, 1.19)	0.99 (0.72, 1.33)	0.95 (0.70, 1.28)	0.95 (0.69, 1.28)
A=2	0.88 (0.64, 1.17)	0.98 (0.69, 1.31)	0.98 (0.67, 1.29)	0.98 (0.66, 1.30)
A=3	0.87 (0.64, 1.13)	0.85 (0.59, 1.17)	0.87 (0.61, 1.15)	0.83 (0.60, 1.14)
A=4	0.78 (0.62, 0.97)	0.85 (0.64, 1.08)	0.84 (0.63, 1.04)	0.85 (0.63, 1.10)
A=5	0.91 (0.75, 1.11)	0.96 (0.73, 1.23)	0.99 (0.73, 1.23)	1.01 (0.73, 1.30)

treatment mechanism that satisfies the ETA assumption, we argue that the results offered by these three estimators must also be treated with great caution. Since, strictly speaking, static causal effects cannot be identified from the observed data if the ETA assumption is violated, it should in fact make sense that the appropriate response to this problem does not lie in turning to approaches that aim to estimate such parameters by relying on untestable modelling assumptions, but rather in adapting the definition of the parameter of interest in a way that makes the parameter identifiable.

This becomes particularly obvious in cases in which static causal effects are not even well-defined. In the context of studying the causal effect of vigorous physical activity on mortality in the elderly, for instance, it makes little sense to talk about the counterfactual outcome distribution we would observe if all subjects were assigned to high levels of activity since serious health problems would prevent a considerable proportion of subjects from complying with such an assignment. Causal effects defined on the basis of realistic individualized treatment rules and intention-to-treat rules address this problem by explicitly taking into account the set of treatment options that are realistically available to each subject. Such effects are therefore well-defined and identifiable even if the full set of treatment options is not available to some subjects. The estimates of such effects reported here suggest that high levels of vigorous physical activity may confer reductions in mortality risk on the order of 15-30%, although in most cases the evidence for such an effect does not quite reach the 0.05 level of significance. Estimates of static causal effects, in contrast, suggest a statistically significant reduction in mortality risk on the order of 50-75%, a finding that given the estimated bias of the IPTW estimator, must be viewed as highly suspect.

A possible extension to the analysis we present here consists of data-adaptively selecting the value for α in definition (3) of the set of realistic treatment options, arbitrarily set by us as $\alpha = 0.05$. For very small values of α, estimators of causal effects based on realistic individualized treatment rules and intention-to-treat rules may still be affected by a practical violation of the ETA assumption. As the value for α is increased, on the other hand, the corresponding causal effects become more and more different from the static causal effect that they are in some sense intended to approximate. A more sophisticated analysis might thus attempt to use the approach introduced by Wang et al. (2006) in order to find the smallest value of α for which the ETA bias of the ITPW estimator is estimated to be negligible. Future research will be required to investigate this approach further.

6 Acknowledgements

We would like to thank Dr. Ira Tager from the Division of Epidemiology at the UC Berkeley School of Public Health for kindly making available the dataset that was used in our data analysis. His work on the SPPARCS project was supported by a grant from the National Institute on Aging (RO1-AG09389).

References

Ainsworth, B., Haskell, W., Leon, A., Jacobs, D. J., Montoye, H., Sallis, J., and Paffenberger, Jr., R. (1993). Compendium of physical activities: classification of energy costs of human physical activities. *Medicine and Science in Sports and Exercise*, 25:71–80.

CDC (1989). Surgeon general's workshop on health promotion and aging: summary recommendations of physical fitness and exercise working group. *Journal of the American Medical Association*, 262:2507–2510.

CDC (1996). Physical activity and health: a report of the surgeon general. Altanta, Georgia: US Department of Health and Human Services, CDC.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

Esposito, K., Pontillo, A., Di Palo, C., Giugliano, G., Masella, M., Marfella, R., and Giugliano, D. (2003). Effect of weight loss and lifestyle changes on vascular inflammatory markers in obese women: a randomized trial. *Journal of the American Medical Association*, 289(14):1799–1804.

Hosmer, D. and Lemeshow, S. (2000). *Applied Logistic Regression*. Wiley, New York.

Hosmer, D. W., Hosmer, T., le Cessie, S., and Lemeshow, S. (1997). A comparison of goodness–of–fit tests for the logistic regression model. *Statistics in Medicine*, 16:965–980.

Hosmer, D. W. and Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics*, A10:1043–1069.

Lee, I., Sesso, H., Oguma, Y., and Paffenberger, Jr., R. (2003). Relative intensity of physical activity and risk of coronary heart disease. *Circulation*, 107(8):1110–116.

Nagi, S. (1976). An epidemiology of disability among adults in the United States. *Milbank Quartely*, 54:439–468.

Neugebauer, R. and van der Laan, M. (2005). Why prefer double robust estimates in causal inference? *Journal of Statistical Planning and Inference*, 129:405–426.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. *Statistical Science*, 5:465–480 (1990).

Robins, J. (1986). A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy survivor effect. *Mathematical Modelling*, 7:1393–1512.

Robins, J. (1987). Addendum to "a new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy survivor effect" [Math. Modelling **7** (1986) 1393-1512]. *Computers and Mathematics with Applications*, 14:923–945.

Robins, J. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association 1999*, pages 6–10.

Rosano, C., Simonsick, E., Harris, T., Kritchevsky, S., Brach, J., Visser, M., Yaffe, K., and Newman, A. (2005). Association between physical and cognitive function in healthy elderly: the health, aging, and body composition study. *Neuroepidemiology*, 24(1-2):8–14.

Rosow, I. and Breslau, N. (1966). A Guttman health scale for the aged. *Journal of Gerontology*, 21:556–559.

Rubin, D. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6:34–58.

Sinisi, S. and van der Laan, M. (2004). Deletion/Addition/Substitution Algorithm in Learning with Applications in Genomics. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 18.

Tager, I., Hollenberg, M., and Satariano, W. (1998). Self-reported leisure-time physical activity and measures of cardiorespiratory fitness in an elderly population. *American Journal of Epidemiology*, 147:921–931.

van Dam, R., Schuit, A., Feskens, E., Seidell, J., and Kromhout, D. (2002). Physical activity and glucose tolerance in elderly men: the Zutpen Elderly study. *Medicine and Science in Sports and Exercise*, 34:1132–1136.

van der Laan, M. and Petersen, M. (2007). Causal effect models for realistic individualized treatment and intention to treat rules. *The International Journal of Biostatistics*, 3(1):Article 3.

van der Laan, M. and Robins, J. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer Series in Statistics. Springer Verlag.

van der Laan, M. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):Article 11.

Wang, Y., Petersen, M., Bangsberg, D., and van der Laan, M. (2006). Diagnosing Bias in the Inverse-Probability-of-Treatment-Weighted Estimator Resulting from Violation of Experimental Treatment Assignment. Technical Report 211, UC Berkeley Division of Biostatistics Working Paper Series.

A Targeted MLE of the causal relative risk

Let $\psi_a = E[Y_{d(a,W)}]$ and consider the parameter

$$\theta = \frac{E[Y_{d(a,W)}]}{E[Y_{d(0,W)}]} = \frac{\psi_a}{\psi_0}. \tag{34}$$

Since we have already derived the efficient influence curve of ψ_a as

$$D^{\psi_a}(P) = \frac{I(A = d(a,W))}{g(A \mid W)}\Big[Y - Q(A,W)\Big] + Q(d(a,W),W) - \psi_a, \tag{35}$$

we can use the δ-method to find the efficient influence curve for θ. Specifically, we have that

$$\theta = f(\psi_a, \psi_0) = \frac{\psi_a}{\psi_0} \tag{36}$$

and

$$Df = (1/\psi_0, -\psi_a/\psi_0^2) \tag{37}$$

so that the efficient influence curve for θ is given by

$$\begin{aligned} D(P) &= Df(D^{\psi_a}(P), D^{\psi_0}(P))^T \\ &= \frac{1}{\psi_0}\left\{\frac{I(A = d(a,W))}{g(A \mid W)}\Big[Y - Q(A,W)\Big] + Q(d(a,W),W) - \psi_a\right\} - \\ &\quad \frac{\psi_a}{\psi_0^2}\left\{\frac{I(A = d(0,W))}{g(A \mid W)}\Big[Y - Q(A,W)\Big] + Q(d(0,W),W) - \psi_0\right\} \\ &= \frac{1}{\psi_0}\Big[I(A = d(a,W)) - \theta I(A = d(0,W))\Big]\frac{Y - Q(A,W)}{g(A \mid W)} + \\ &\quad \frac{1}{\psi_0}\Big[Q(d(a,W),W) - \theta Q(d(0,W),W)\Big]. \end{aligned} \tag{38}$$

The component of this influence curve corresponding to the score for $P(Y \mid A, W)$ is given by

$$D(P) - E[D(P) \mid A, W] = \frac{1}{\psi_0}\Big[I(A = d(a, W)) - \theta I(A = d(0, W))\Big]\frac{Y - Q(A, W)}{g(A \mid W)}. \tag{39}$$

The covariate $h(P_n^0)(A, W)$ needed for the univariate regression to update the initial fit for Q is thus given by

$$\begin{aligned} h(P_n^0)(A, W) &= \frac{I(A = d(a, W)) - \theta I(A = d(0, W))}{g(P_n^0)(A \mid W)\psi_0} \\ &= \frac{I(A = d(a, W)) - \psi_a/\psi_0 I(A = d(0, W))}{g(P_n^0)(A \mid W)\psi_0}. \end{aligned} \tag{40}$$

To obtain a feasible $h(P_n^0)(A, W)$, we substitute

$$\psi_{a,n} = \frac{1}{n}\sum_{i=1}^{n} Q(P_n^0)(d(a, W_i), W_i) \tag{41}$$

and

$$\psi_{0,n} = \frac{1}{n}\sum_{i=1}^{n} Q(P_n^0)(d(0, W_i), W_i) \tag{42}$$

for ψ_a and ψ_0, respectively. Let ϵ_n denote the MLE of ϵ in $Q(P_n^0)(\epsilon)$ and let

$$Q_n^1(A, W) = \frac{1}{1 + \exp(-m_n^0(W) - \epsilon_n h(P_n^0)(A, W))}. \tag{43}$$

Iterate this process k times until ϵ_n has become sufficiently small. Then the targeted MLE of θ is given by

$$\theta_n^{tMLE} = \frac{\sum_{i=1}^{n} Q_n^k(d(a, W_i), W_i)}{\sum_{i=1}^{n} Q_n^k(d(0, W_i), W_i)}. \tag{44}$$

The covariate $h(P_n^0)(A, W)$ for the corresponding intention-to-treat relative risk parameter can similarly be derived as

$$\begin{aligned} h(P_n^0)(A, W) &= I(a \in \mathcal{D})\left[\frac{1}{\psi_0} - \frac{\psi_a}{\psi_0^2}\right] + \\ &\quad I(a \notin \mathcal{D})\left[\frac{I(A = d(a, W)) - \psi_a/\psi_0 I(A = d(0, W))}{g_n^0(A \mid W)\psi_0}\right]. \end{aligned} \tag{45}$$

Chapter 7

Biomarker Discovery

7.1 *Targeted Methods for Biomarker Discovery, the Search for a Standard*

The following article appears as it was originally published on the University of California, Berkeley Division of Biostatistics Working Paper Series website in 2008, http://www.bepress.com/ucbbiostat/paper233/.

Targeted Methods for Biomarker Discovery
The Search for a Standard

Catherine Tuglus [1*], Mark J. van der Laan [1,2]

[1] *Division of Biostatistics, School of Public Health, University of California Berkeley*

[2] *Department of Statistics, University of California Berkeley*

Abstract

Often biomarker analyses involve numerous variables with complicated and generally unknown correlation structure, and biomarker detection methods aimed at identifying causally related biomarkers often result in measures that are difficult to interpret and nearly impossible to compare across studies. In response to the FDA call for method regulation, we present targeted variable importance (tVIM) as a standardized method for biomarker discovery. Based on targeted maximum likelihood theory, these estimates are robust to model misspecification and, under specified conditions, interpretable as a causal effect, making them reproducible across populations. In simulation, we compare tVIM to univariate linear regression, LASSO penalized multiple regression, and randomForest under conditions of increasing correlation. Based on ranked variable lists, tVIM coupled with a data-adaptive model selection method is more resilient to increases in correlation, outperforming the other methods. In application we apply the tVIM to the van't Veer breast cancer data. Overall, tVIM appears to rank biologically relevant genes at the top its list. Given extreme correlations, methods to reduce bias and provide realistic gene lists are also discussed.

*Correspondence to: 101 Haviland, MC 7358, Berkeley, CA 94720; email: ctuglus@berkeley.edu

1. INTRODUCTION

The use of biomarkers in disease diagnosis and treatment has grown rapidly in recent years, as microarray and sequencing technologies capable of detecting biological signatures have become more effective and efficient research tools. In an attempt to create a level of quality assurance with respect to biological and more specifically biomarker research, the FDA has called for the development of a standard protocol for biomarker qualification [1]. Such a protocol would define "evidentiary" standards for biomarker usage in areas of drug development and disease treatment and provide a standardized assessment of a biomarker's significance and biological interpretation. This is especially relevant in clinical trials, where the protocol would prohibit the use of unauthenticated biomarkers to determine treatment regime, resulting in safer and more reliable treatment decisions [2]. Consequentially identifying accurate and flexible analysis tools to assess biomarker importance is essential.

Many biomarker discovery methods only measure the association between the marker and the biological outcome. However a significant association is often difficult to interpret and does not guarantee that the biomarker will be a suitable and reliable drug candidate or diagnostic surrogate. This is especially true with genomic data, where genes are often present in multiple pathways and can be highly correlated amongst themselves. Applying association-based methods to this data will often lead to a long and ambiguous listing of biomarkers, which can be expensive to analyze.

Ideally, biomarker discovery analyses want to identify markers that systematically effect the outcome through a biological pathway or mechanism, in other words markers causally related to the outcome of interest. Once these markers are identified, they can be further analyzed and eventually applied as potential drug targets or prognostic markers. Due to the complex nature of the human genome, this is not a straight forward task and certain assumptions are required to identify a causal effect.

A causal effect is often thought of in terms of an intervention on a causal diagram [3]. If we define a causal diagram with nodes, $\{A, W, Y\}$, and conditional relationships $A|W$ and

$Y|A, W$, we can observe the causal effect of A on Y controlling for W through intervention, setting $A = a$ for multiple values a and observing the response $Y = y_a$. In terms of a clinical trial this is the equivalent to observing the response (Y) to an assigned treatment ($A = a$).

Causality is only identifiable if A occurs prior to or concurrent with response Y and all nodes in the path between A and Y are included in the diagram. This is equivalent to requiring that there are no unmeasured confounders, or that conditional on measured W the assignment of A is independent of Y (e.g. randomized trial). For example, often the parameter β estimated from univariate regression $E[Y|A] = \beta A$ is interpreted as a measure of importance for A. However, by ignoring potential confounders (W) (i.e. other genes in the pathway or system), this approach estimates a measure only interpretable as a causal effect in the case where there are no confounding factors (all W is kept constant at all values of A). Consequently, it will often identify genes as important which are merely correlated with causally related variables.

Identifiability issues can also arise for causal effects when for a given set W, some levels (i.e. treatments) a for A are unlikely or not possible in the current study population. This often arises when confounders W are highly correlated with the variable A. Consider a drug trial for breast cancer. If the drug is assigned to people who also undergo radiation treatment and individuals in the placebo group generally do not undergo radiation, then the effect of the drug and radiation on the cancer cannot be distinguished. The assumption that all levels of A are possible for any given levels of W is sometimes referred to as the experimental treatment assumption (ETA) [4].

Multivariate methods seek to estimate the causal effect by controlling for confounders within a model. If the parameter of interest and all confounders are correctly accounted for in the model, and the stated assumptions hold, this method can produce an estimate of the causal effect. Though in reality, the model is generally not correct, and high correlation among the variables can lead to bias in the estimate. In the case of high dimensional data commonly found in biomarker studies, the number of covariates is larger than the number of observations making model selection a necessity.

Standard model selection methods often over-fit the data and can end up removing the

variable of interest. More advanced methods, such as penalized regression often combat over-fitting by using cross-validation. In particular, LASSO penalized regression uses cross-validation to determine the amount of shrinkage for its coefficients, resulting in a high dimensional fit where all variables have coefficient (importance) values.

Supervised learning methods, such as randomForest, are also commonly used in biomarker discovery studies. RandomForest is a tree-based regression algorithm that exploits boosting to reduce the variance of a bias predictor fit. Implementation of randomForest in R (randomForest()) provides two measures of importance based on the sensitivity of the 'out-of-bag" error rate and node classification under perturbation of the value of the variable [5]. Though, these measures lack causal interpretation, they are often used in practice when the data is high dimensional and over-fitting is a risk.

Both lasso and randomForest lack formal inference and depend on additional boosting for inference. This is generally computational prohibitive in biomarker discovery analyses where the number of variables is quite large.

In general, causal effects are often difficult if not impossible to estimate correctly, especially in high-dimensional and highly correlated genomic data. The specific assumptions they require (randomized treatment, experimental treatment assignment, etc.) are often only fully realized in randomized trials, making their utility in a standard protocol limited. However, measures which are causally interpretable in randomized trials, can still be biologically interpretable in observation data as measures of importance.

Here, we present the typical representation of the direct causal effect as a potential measure of biomarker importance

$$\Psi(P) = \mathbb{E}[\mathbb{E}_{\mathbb{W}}[Y|A = a, W] - \mathbb{E}[Y|A = 0, W]]$$

Given the observed data $O = (A, W, Y) \sim P$, this measure corresponds to the effect of a biomarker (A) on the outcome (Y), adjusting for confounders (W). Here, A can represent a single biomarker or set of biomarkers. This article will focus on the univariate case.

This measure can be estimated robust to model mis-specification and with formal inference using targeted maximum likelihood estimation [6]. Targeted Maximum Likelihood (tMLE) methodology reduces the bias for the targeted parameter by maximizing the likelihood in a direction which corresponds to the best estimate of the targeted parameter [6]. Consequently we will refer to this estimate as the W-adjusted targeted variable importance (tVIM).

A major benefit of the tVIM measure is that it is a biologically interpretable measure. In the case of a randomized trail, tVIM provides an estimate of the causal effect of A on the outcome Y, where W would contain all variables which confound the effect of A on Y. In the case of observational data, one can still interpret $\Psi(P)$ as estimated causal effect that would have been observed under experimental conditions which only control for the given set variables, W.

Additionally tVIM can be adapted to address ETA violations. When W is highly correlated with A, we can use tVIM in conjunction with a correlation cut-off to provide a realistic ranking of variables. Given a correlation cut-off, tVIM will identify causally related variables as well as all variables the data is unable to disentangle due to the high correlation structure. The optimal cut-off decreases bias in the estimate while still maintaining the optimal level of reproducibility. Targeted Variable Importance measures also have formal inference and multiple testing methods based on the influence curve which allows estimation of the overall joint distribution without resampling [7].

Variable importance using tMLE methods has been previously presented in [8] for a binary A. In this article, we present the semi-parametric version of variable importance, which is flexible enough to accommodate the wide variety of biomarker data types (continuous, binary, etc.). Though, we will primarily explore biomarker discovery with respect to gene expression data (continuous A). This measure of variable importance was first presented in [9], and estimated using tMLE in [6].

In this article we present tVIM as a candidate standardized measure of biomarker importance. We demonstrate its efficacy and functionality through both simulation and application. Simulations provide a performance assessment of tVIM under increasing levels

of correlation. We demonstrate the accuracy in which tVIM can detect "true" variables from amongst increasingly correlated "decoy" variables. Additionally we also evaluate the accuracy of three commonly used methods for biomarker discovery under the same conditions, univariate linear regression, lasso regression, and randomForest. We assess and compare these methods based on their ROC curves (a representation of Sensitivity and Specificity) and the length of list required to detect all "true" variables (a representation of Type I error). Methods are applied using the current R version of lm(), lars() from library *lars* [10], and randomForest from the library *randomForest* [11]. These versions and their implementation of the methods are representative of the current tools available to biologists.

After introducing tVIM more formally and presenting the basics of tMLE, the simulation study is presented. It is followed by a discussion of ETA bias which is problematic for all methods when the data is highly correlated. We then present an application of tVIM to the van't Veer et al. 2002 breast cancer dataset [12]. The van't Veer study is focused on predicting a patient's response to treatment given their gene expression profile. We use tVIM to determine which genes are relevant to treatment response, providing an accurate list of input variables for any prediction algorithm. This is followed by an overall discussion.

2. Targeted Variable Importance

The proposed standard measure of importance, tVIM, is a marginal variable importance measure and is analogous to the variable importance (VIM) measure in [13] which was presented for a binary A. In order to accommodate a more general A (i.e. continuous), tVIM is based on a semi-parametric model approach. This method was first presented in van der Laan 2005 [9], and models of this type have also been considered previously in the literature [14, 15, 16].

Given the observed data defined as $O \sim (A, W, Y)$, where A is the variable (i.e. gene) of interest, W is the set of potential covariables (i.e. genes), and Y is the outcome of interest, we can define a general semi-parametric model as follows

$$\mathbb{E}_P[Y|A,W] = m(A,W|\beta) + g(W)$$

where m is user specified given $m(A=0,W|\beta)=0$ for all β and W, and $g(W)$ is an unspecified function of potential covariates W.

Using this model form for $\mathbb{E}_P[Y|A,W]$, we can represent the difference as

$$\begin{aligned}\mathbb{E}_P[Y|A=a,W] - \mathbb{E}_P[Y|A=0,W] &= m(a,W|\beta) + g(W) - m(0,W|\beta) - g(W) \\ &= m(a,W|\beta)\end{aligned}$$

and can define generally, the tVIM of a particular A on outcome Y controlling for confounders W as

$$\mu(a) = \mathbb{E}_W[m(a,W|\beta)]$$

This is referred to as the W-adjusted variable importance

Given a linear model for $m(A,W|\beta)$ in terms of A (i.e. $(m(A,W|\beta) = AW\beta)$), the importance can be represented as the linear curve, $\mathbb{E}_W[m(A=a,W|\beta)] = a\beta_W\mathbb{E}[W]$, and the tVIM becomes a simple linear combination c^TA. Formal inference can be estimate by applying the delta method. Further detail is provided in section 3 and Appendix I.

Compared to the tVIM method presented in Bembom and van der Laan 2008 [8], this model based approach to variable importance not only accommodates a continuous A but can also incorporate effect modification. For instance, effect modification of A by W_1 can be obtained with the following model

$$m(A,W|\beta) = \beta A + \beta_1 AW_1$$

This is especially relevant in clinical trials where the research is interested in finding genes (i.e. W_1) which modify the causal effect of a given a particular treatment (A) on overall disease response. Another benefit of tVIM for general A is the exclusion of inverse weighting making

this measure more robust to experimental treatment assumption violations, which will be discussed further in section 5.

In this paper we focus on the simplest linear case $m(A, W|\beta) = A\beta$, where the marginal importance of A can be represented by single coefficient value β. This allows us to directly compare with alternative measures of importance obtained from univariate and multivariate regression methods. In this analysis, A is a single biomarker, however the method can be extended to analyze a set of biomarkers $\{A\}$.

3. Targeted Maximum Likelihood for tVIM

Although a "plug-and-chug" estimate of $\mu(a)$ could be achieved using a maximum likelihood estimate of $E[Y|A, W]$ (for instance from linear regression or LASSO), the estimate of $\mu(a)$ would be unnecessarily bias. This is because the Maximum Likelihood estimate is based on the bias-variance trade-off for estimating $\mathbb{E}[Y|A, W]$, not your parameter of interest. Targeted Maximum Likelihood (tMLE) methodology reduces the bias for the targeted parameter by maximizing the likelihood in a direction which corresponds to the best estimate of the targeted parameter [6], resulting in the doubly robust locally efficient estimate.

Targeted Maximum Likelihood updates an initial regression estimate $\mathbb{E}[Y|A, W]$ in a direction which targets the parameter of interest. The update is completed by regressing the outcome on a clever covariate and setting the initial estimate of $\mathbb{E}[Y|A, W]$ as an offset. The clever covariate is determined from tMLE methodology and its derivation can be found in van der Laan and Rubin 2006 [6] and is outlined in Appendix I.

The update is a function of $\mathbb{E}[A|W]$, which is often referred to as the "treatment mechanism." The "treatment mechanism" is an estimate of the effect of confounders, W, on "treatment" (or variable) A. Given correct model specification for either $\mathbb{E}[Y|A, W]$ or $\mathbb{E}[A|W]$, the tVIM estimate is a consistent and asymptotically normal and linear estimate. The estimate is efficient when both models are correctly specified (a.k.a. "locally efficient"). This feature is referred to as "doubly robust." Improving the estimates of $\mathbb{E}[Y|A, W]$ and $\mathbb{E}[A|W]$ using data-adaptive or

super learning algorithms will improve the overall consistency and efficiency of the estimate. The derivation of the targeted MLE methodology for tVIM is summarized in Appendix I and described in further detail in Tuglus and van der Laan 2008 [17].

In biomarker discovery analyses tVIM is applied to all variables within the data matrix W, where W is a matrix of genes, SNPs, or other biological variables of interest. The method is outlined below for a single A, with the possible covariate set W.

There are three initial components necessary for applying targeted Maximum Likelihood methodology to estimate tVIM.

1. A model $m(A, W|\beta)$ satisfying $m(0, W \mid \beta) = 0$ for all β and W. In this case it is defined as $m(A, W|\beta) = \beta A$
2. An initial regression estimate for $Q(A, W) = \mathbb{E}[Y|A, W]$ of the form $\mathbb{E}[Y|A, W] = m(A, W|\beta) + g(W)$, where $g(W)$ is estimated data-adaptively. We recommend using polymars [27, 28], lars [10, 18], or DSA [25]
3. An estimate of the "treatment mechanism" $G(W) = \mathbb{E}[A|W]$, estimated data-adaptively.

Given these three components, tMLE can easily be applied in the following steps

1. Estimate the "clever covariate" which will allow us to update the initial regression in a direction which targets the parameter of interest. In this case the clever covariate is defined as:
$$r(A, W) = \frac{d}{dB} m(A, W|\beta) - \mathbb{E}[\frac{d}{dB} m(A, W|\beta)|W]$$
which for this particular $m(A, W|\beta) = \beta A$ simplifies to $r(A, W) = A - \mathbb{E}[A|W]$
2. Compute the fitted values for your initial estimate of $Q_n^0(A, W)$
3. Project Y onto $r(A, W)$ with $offset = Q_n^0(A, W)$ and define the resulting coefficient as ϵ. This is done using standard software (lm() in R) setting the offset, and projecting onto the model $Y \sim \epsilon' r(A, W) + offset$. Note there is no intercept in your model, only the offset value.
4. Update the initial estimate $\beta_n^0 = \beta_n^0 + \epsilon$ and overall density $Q_n^1(A, W) = Q_n^0(A, W) +$

$\epsilon r(A,W)$. These are now your single-step targeted estimates. Since this is a simple linear model, the single step solution is the final solution.

5. Obtain standard error and inference for β using the empirical estimate of the conservative influence curve. For the true parameter and tMLE updated density, β_0 and Q_0^1, the empirical influence curve for a given A is defined as

$$\hat{IC}(O) = c^{-1}D(O|\beta_0, Q_0^1)$$

with scale factor $c = \mathbb{E}[\frac{d}{d\beta}D(O|\beta_0, Q_0^1)]$ where

$$D_h(p_0)(O) \equiv r(A,W)(Y - m(A,W|\beta_0) - Q_0(0,W))$$

The covariance of β_0 is asymptotically equivalent to the covariance of $IC(O)$. Therefore the empirical estimate of the covariance for parameter estimate β_n is

$$\Sigma_n = \frac{1}{n}\sum \hat{IC}(O)\hat{IC}(O)^T$$

such that

$$\sqrt{n}(\beta_n - \beta_0) \sim N(0, \Sigma_n)$$

Covariance can also be estimated by bootstrap estimates of β, but this would requiring extra computational time. If $\mathbb{E}[A \mid W]$ is estimated consistently, then the variance estimates based on the influence curve are consistent or asymptotically conservative. See Tuglus and van der Laan 2008 [17] and van der Laan and Robins 2003 [26] for supporting theory and formal proof.

6. Using the estimated covariance, test the hypothesis $H_0 : \beta_n(j) = 0$, using a standard test statistic to obtain p-values.

$$T_n(j) = \frac{\sqrt{n}\beta_n}{\sqrt{\Sigma_n(j,j)}} \underset{n\to\infty}{\sim} Normal(0,1)$$

Also note that inference for linear combinations can be obtained by applying the delta method [17] (see Appendix I)

4. Simulations

In this paper we compare tVIM to three other methods commonly used for determining variable importance in biomarker discovery analyses: univariate linear regression [18], LASSO regression with cross-validation based model-selection - using R package *lars* [10], and randomForest [5] - using R package *randomForest* [11]. Importance measures for univariate linear regression and LASSO regression are represented by the associated coefficient value. RandomForest provides two measure of importance based on the effect perturbing the variable of interest has on overall classification error and node splits. Each is summarized briefly below.

Note that given any estimate, bootstrap sampling may be used to provide standard error estimates and p-values. However in this analysis we choose to compare the methods based on their current merits and accessible implementation, not on any additional processing. Also, in biomarker discovery there are thousands of genes and bootstrap sampling is computationally expensive and impractical.

Univariate Linear Regression (LM): Marginal variable importance is represented by the coefficient and p-value resulting from the univariate linear regression fit, $\mathbb{E}[Y|A] = \beta A$. This method does not account for any confounding and will often misclassify genes correlated with the "true" genes as significant. In most situations this importance measure can not be interpreted in as a causal effect.

Penalized Regression - LASSO (Q): Marginal Variable Importance is represented by the coefficient of A in LASSO main term fit of $Q(A, W_s) = \mathbb{E}[Y|A, W_s]$, where $W_s \subset W$ representing the subset of W found significant according to their univariate regression on Y. LASSO is applied using R package *lars* [10], which does not provide any formal inference therefore p-values are not recorded. Results are compared based on the variable

importance measure and its rank. LASSO does attempt to account for confounding, but will only allow for n-1 non-zero coefficient values, making its applicability to high dimensional data limited [19]. LASSO is also maximum likelihood method which focuses on estimating the overall distribution $\mathbb{E}[Y|A,W]$ and not the parameter of interest.

Targeted Variable Importance (tVIM): Marginal Variable Importance measure is obtained from applying targeted MLE to the initial density estimate provided by LASSO fit $Q(A,W_s)$. Coefficient of A is targeted directly, and p-values are provided based in the covariance estimate of the conservative empirical influence curve. The measure will be represented and compared in terms of the coefficient β as presented in section 2.

randomForest (RF1 and RF2): Two measures of importance, RF1 and RF2, are provided by the R function randomForest(). The function is applied directly to the full data matrix W using R package *randomForest* [11], using the default setting with 500 trees.

- RF1: RandomForest importance measure based on "out-of-bag" error rate [5, 11] (no p-values provided)
- RF2: RandomForest importance measure based on accuracy of node split [5, 11] (no p-values provided)

Though it does not estimate the same measure as LM, LASSO, or tVIM, randomForest (RF) is a tree-based algorithm developed by Breiman 2001 [5] commonly used in biomarker discovery analyses. However due to the nature of randomForest, there is no guarantee that all biomarkers will receive a measure of importance. Also no formal inference is available; therefore no p-values are recorded.

P-values for LM and tVIM estimates are calculated using a standard t-test and are subjected to the Benjamini & Hochberg step-up FDR controlling procedure [20] to control for multiple testing.

We compare methods based on their ability to produce an accurately ranked list of genes. Often in practice, biomarkers are ranked in the order of increasing p-value, where markers with

p-value below a particular cut-off are defined as "important." Alternatively if no p-values are provided, the biomarkers may be ranked by their importance measure, and the cut-off would be based on a numeric threshold or required level of importance. Inaccuracies in these lists and rankings often arise when the data is highly correlated. Therefore, we will evaluate each methods ability to produce an accurate ranking under increasing levels of correlation.

We simulate data to compare the four approaches under increasing correlation levels using a diagonal block correlation structure. The structure of the simulated data allows us to study the effects that both correlated and uncorrelated variables have on the reported importance of the true variables. For each approach, the biomarkers will be ranked by the resulting importance measure and p-value (when available). The sensitivity and specificity of methods will be compared based on both p-value and rank-based cut-off values, and will be summarized using ROC plots. We will determine the ability of each approach to identify the true variables and each variables true importance rank by comparing the length of list required to label all true variables as "important."

4.1. Simulated Data

The full data is defined as $O = (W, Y) \sim P$, with covariate matrix W and outcome Y. Covariate matrix W consists of J=100 variables with n=300 observations simulated from a multivariate normal distribution with block diagonal correlation structure and mean vector created by randomly sampling mean values from $\{0.1, 0.2, ..., 9.9, 10.0, 10.1, ..., 50\}$, resulting in K=10 independent sets of variables, each correlated according to an exchangeable correlation structure with variance=1 and specified correlation ρ_{TRUE}. This forms a J by n matrix where each set of ten is correlated among themselves but independent from all other variables.

Outcome Y is simulated from a main effect linear model using one variable from each of the K sets. These K variables are designated as "true variables." The importance of a variable is determined by its coefficient value in simulation. Two sets of values are used: a constant value ($\{\beta_k = 4 : k = 1, ..., 10\}$) and an increasing set ($\{\beta_k = k : k = 1, ..., 10\}$). A normal error with mean zero and variance σ_Y is added as noise.

Simulations are run for $\rho_{TRUE} = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$ and $\sigma_Y = 1, 10, 20$ using both sets of coefficient values. At $\sigma_Y = 1$ all methods perform very well, resulting in p-values much below zero. At $\sigma_Y = 20$ all methods became largely erratic and overcome by noise. Simulations at $\sigma_Y = 10$ had enough variation to highlight the different strengths of each method and are considered the most realistic noise scenario. For these reasons, only $\sigma_Y = 10$ results are presented in full.

4.2. Methods

For clarity, we define the full set of J variables (i.e. biomarkers) as W^*, where A can be viewed as a single variable $A = W_j^*$ in W^*, and the remaining variables are defined as the covariate set $W = W_{-j}^*$, for all $j = 1 \ldots J$. Importance measures according to the five methods outlined previously (LM, LASSO(Q), tVIM, RF1, and RF2) are calculated for each individual variable (i.e. biomarker), A.

We first apply univariate regression to all J biomarkers, estimating $E[Y|A] = \beta_A^{LM} A$. We record each β_A^{LM}, as the LM importance measure along with its associated p-value, and adjust for multiple testing using Benjamini & Hochberg step-up FDR controlling procedure [20] applied using the mt.rawp2adjp() R function in package *multtest* [21].

To facilitate estimation in LASSO, we first reduce the possible covariate set W, to only those variables which are univariate significant with marginal LM adjusted p-value less than $\alpha = 0.05$. We define this reduced set for a given A as W_s, and apply LASSO penalized regression to the covariate set $\{A, W_s\}$, giving us an initial estimate $Q(A, W_s) = \mathbb{E}[Y|A, W_s]$. The coefficient of A from the LASSO fit is recorded as the LASSO (Q) importance measure, and this fit is then used as the initial estimate for tMLE. We use the R library *lars* implementation of LASSO [10, 18], which does not provide formal inference therefore p-values are not recorded.

We estimate $G(W) = \mathbb{E}[A|W]$ using LASSO as well citing that the additive main effect form of a LASSO derived model accurately reflects the correlation structure of the data giving us a correct estimate of $G(W)$. This guarantees under minimal ETA violations that we will obtain a consistent estimate due to the double robust nature of the tVIM measure [6]. We record the

updated tVIM measure as well as its respective p-values. All p-values are adjusted for multiple testing using the Benjamini - Hochberg step-up FDR controlling procedure [20].

RandomForest is applied directly to the full data W, and importance measures RF1 and RF2 are calculated internally. Importance measures for randomForest cannot be directly compared because they are not on the same scale as LM, LASSO (Q), or tVIM estimate. Instead we compare based on importance rank.

4.3. Results

For each $\{\rho, \sigma_Y\}$ set, simulations of 100 are completed. Recorded importance measures and p-values are translated into a list of ranks, and the ranks are averaged over the 100 iterations. A rank of one being the largest importance value or smallest p-value. Sensitivity and Specificity calculations for each simulation are also determined for each individual iteration and averaged across the 100 iterations to produce the final estimates.

Simulation results are summarized here in terms of Area Under the Curve (AUC) and Length of List. Additional measures of performance (Type I error, Power, accuracy of importance rank and value) can be found in Appendix I and in the original technical report [17].

Analysis found no appreciable difference when ranking by measure or p-value for LM and tVIM in these simulations, therefore results in terms of measure will be presently in more detail allowing us to include LASSO (Q) and RandomForest measures in all comparisons.

4.3.1. Area Under the Curve (AUC) The simulations are set up to test the ability of each method to detect (or classify) the true important variables. The overall performance of a classifier is often summarized in terms of the AUC, the Area Under the Curve derived from the basic ROC curve, which plots the true positive rate (Sensitivity) by the false positive rate (1-Specificity) [22]. Under pure noise conditions $AUC = 0.5$, indicating that at any threshold the false positive and true positive rate are equal (random classifier). The more convex the curve becomes, the higher the AUC, and the better the classifier, and a perfect classifier will have AUC=1. Here, we use the R function AUCi() from R package *ROC* which uses integrate()

to calculate the AUC [23]. The calculated AUC values are plotted versus correlation for each of the five methods using importance measure importance rank, and p-values when available for correlations, $\rho_{TRUE} = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$ (Figure 1).

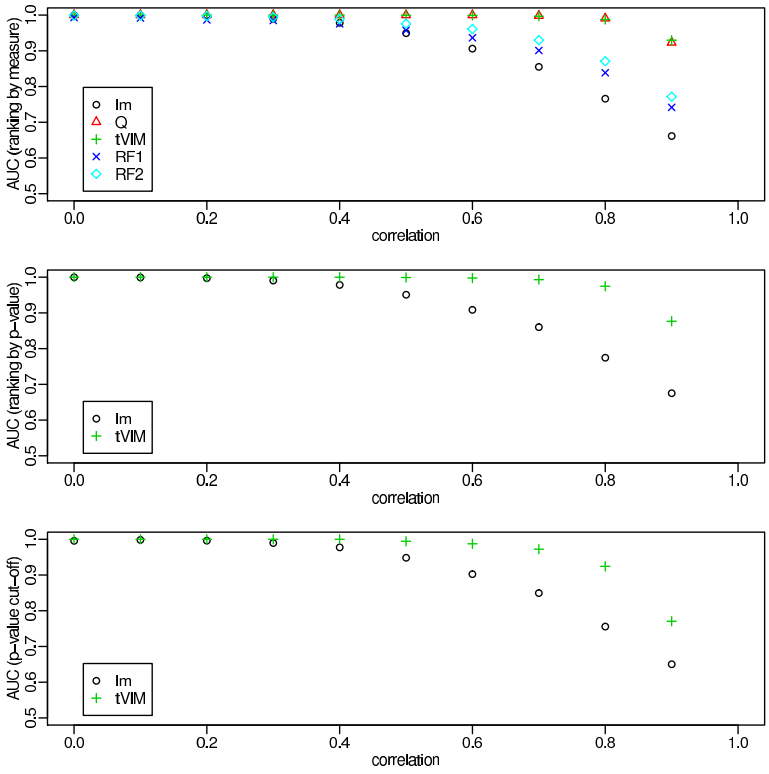


Figure 1: AUC value from ROC curves by $\rho = 0, ..., .9$ completed for (top) ranking by measure (middle) ranking by p-value, and (bottom) p-value cut-off. The later two only contain values for linear regression and tVIM ($\sigma_Y = 10$)Note: minimum AUC is 0.5, maximum and optimum is AUC=1. Simulation is done with $\sigma_Y = 10$ for n=300 with total number of variables at 100 of which 10 are truly related to the outcome. At zero correlation, LASSO (Q), tVIM, and LM perform perfectly with AUC=1. Plots are shown for constant $\beta = 4$, but results are comparable when $\beta = \{1, ..., 10\}$.

From Figure 1, we can see that tVIM performs well up to $\rho = 0.6$, performing only marginally better than Q for $\rho > 0.2$, but with AUC visibly greater than randomForest and LM as correlation increases. As expected LM is most susceptible to increases in correlation, performing perfectly when correlation is zero, but falling consistently as correlation increases, reaching below 0.8 by $\rho = 0.5$.

4.3.2. Average Length of List We can also compare the methods based on the average length of list required to detect all "true" variables. Having a short and accurate list allows the biologist to spend money analyzing the top genes with confidence, knowing that the most important genes are at the top of the list.

The average required list length to find all 10 "true" variables is plotted versus correlation for all five measures and two p-value average ranked lists. These plots are shown for both constant $\beta_{true} = 4$, and $\beta_{true} = \{1 \dots 10\}$. The more detailed required length of list for $k = 1, ..., 10$ true variables for each available ranked list (rank by measure, rank by p-value) at each correlation level as well as plots of the average rank and importance value can be found in Tuglus and van der Laan 2008 [17] and Appendix II.

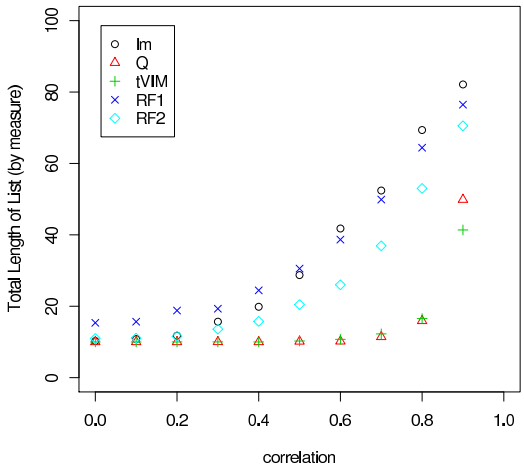


Figure 2: Total length of list required to have all ten true variables in the list by $\rho = 0, ..., .9$, ranking by importance measure. ($\sigma_Y = 10$) Results for univariate regression (LM), LASSO (Q), targeted Variable Importance with LASSO (tVIM) and two randomForest based importance measures (RF1, RF2) are shown. Here β_{TRUE} is constant at 4.

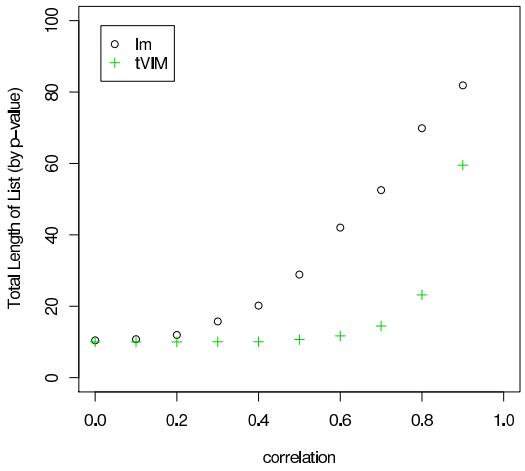


Figure 3: Total length of list required to get all ten true variables by $\rho = 0, ..., .9$, ranking by p-value. ($\sigma_Y = 10$) Results for univariate regression (LM), and targeted Variable Importance with LASSO (tVIM) are shown. Here β_{TRUE} is constant at 4.

Length of list is a direct reflection of Type I error or false discovery rate. We see that overall tVIM performs well up to correlations of 0.9, though the improvement over LASSO is less clear when β_{TRUE} is constant (Figures 2, 3). In the case where $\beta_{TRUE} = \{1, ..., 10\}$ (Figures 4, 5), the improvement of tVIM over LASSO is more pronounced, but detection of the first

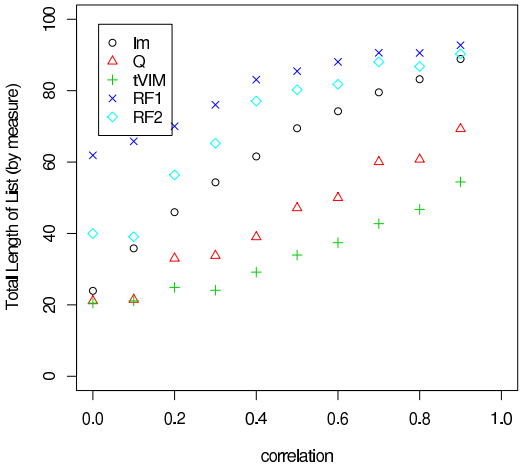


Figure 4: Total length of list required to have all ten true variables in the list by $\rho = 0, ..., .9$, ranking by importance measure. ($\sigma_Y = 10$) Results for univariate regression (LM), LASSO (Q), targeted Variable Importance with LASSO (tVIM) and two randomForest based importance measures (RF1, RF2) are shown. Here β_{TRUE} is set at $\{1, ..., 10\}$.

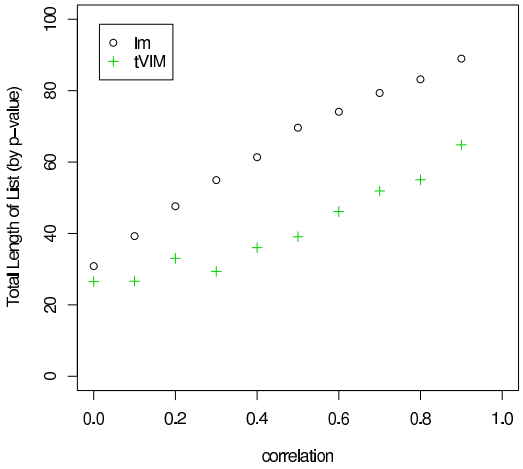


Figure 5: Total length of list required to get all ten true variables by $\rho = 0, ..., .9$, ranking by p-value. ($\sigma_Y = 10$) Results for univariate regression (LM), and targeted Variable Importance with LASSO (tVIM) are shown. Here β_{TRUE} is set at $\{1, ..., 10\}$.

variable (with the lowest β value) is difficult for all methods. When ranking by measure or p-value, all methods have their lowest list length around 20 variables while the total number of variables expected is 10. In contrast, when β was constant at value 4, the lowest list length was near its minimum at 10 (Figures 2, 3). The shift in list length most likely due to the importance value for the variable associated with $\beta = 1$. At such a high noise level ($\sigma_Y = 10$), the lower importance values are more difficult to distinguish from the noise. This is apparent by comparing the average importance rank and average importance value for the variable with $\beta = 1$ (see Appendix II). The rank is much higher than 10, but the value is close to one as it should be.

In general, tVIM has the shortest list and is less affected than any other methods by increases in correlation. Increases above the minimal list length were seen for both random forest and univariate regression methods at correlation values greater than 0.4, with random forest faring better at higher correlations. When ranking by p-value and similar trend for univariate regression was apparent.

4.4. Discussion - Simulations

These simulations address each methods ability to accurately identify the causally related genes as the correlation among variables increases. Though tVIM performs better than the three other methods, it is still sensitive to more extreme correlations (0.7-0.9). Our simulations show only a small increase in bias for the measure of the true variables at higher correlations (see Appendix II). However, in practice, high correlation can adversely effect the tVIM estimate due to violation of the experimental treatment assumption. The increased length of the variable list when ranked by importance measure at correlation 0.8 and 0.9 indicates that tVIM cannot distinguish the true variable from among a group of variables when correlation is very high.

5. Experimental Treatment Assumptions and Consequences of its Violation

When variables are highly correlated with the variable of interest A, ETA violations often occur, which reduces the ability to estimate the effect accurately. Formally, the Experimental Treatment Assumption (ETA) states that the probably of A given W must always be positive for all possible sets (a, W), $(P(A|W) > 0 \ \forall(a, W))$ [4]. In other words all values of A must be possible given any observed set of values W, and no W can be a perfect predictor of A. If either is invalid, estimation of the effect of A will require extrapolation. This introduces bias and, if the ETA violation is extreme enough, can result in a non-identifiable importance estimate.

If our semi-parametric model is correct extrapolation is less of a concern. However in practice, we cannot assume a correct semi-parametric model. In the more realistic case, where we view our importance parameter as a projection onto a working semi-parametric model, violations of ETA can result in a highly sensitive estimate of $Q(A, W)$ leading to instability in the importance (parameter) estimate.

Variable importance measures are also effected by ETA violations through the form of the empirical influence curve used in targeted Maximum Likelihood Estimation (see Appendix

I). The methods for binary A presented in Bembom et al 2009 [8] use inverse weights of the treatment mechanism ($\frac{1}{P(A=a|W)}$) for the tMLE update and inference calculation. When the $P(A = a|W)$ becomes very small from ETA violation, these weights explode leading to unreliable importance estimates. In comparison, the effect of ETA violation on the semi-parametric variable importance presented here is less extreme, but still a concern. It's influence curve is weighted by $(A - E[A|W])$ (for the univariate case), which effectively downplays observations responsible for ETA violations (see Appendix I). Under large ETA violation, the measure is only accounting for a small subset of the observations making it a less applicable and interesting importance.

ETA violations can often be avoided if the "problem" variables (the variables highly correlated with the gene of interest A), are removed from the set of confounders (W). One simple method is to apply a correlation cut-off, where all W whose correlation with A is greater than a particular correlation (ρ_δ), are removed from the set of possible confounders for variable A prior to the application of tVIM method. We explored this briefly through simulation.

In simulation study analogous to the previous set-up, a correlation cut-off was applied to subset W_s for each A before LASSO analysis. In this scenario, W_s is restricted to all $W_i \in W_s$ where $cor(W_j, W_i) < \rho_\delta$, for various cut-offs $\rho_\delta = \{0.5, 0.75, 0.9, 1\}$. We applied this method to our simulated datasets from the previous section. Results showed that such a restriction resulted in the elimination of relevant W_i from the estimate of $\mathbb{E}[Y|A, W_s]$. In other words when A_d is a decoy variable highly correlated with a true variable W_t. Restrictions on the covariate set remove W_t from the possible covariate set for A_d, resulting in A_d having a higher and more significant importance that it would have otherwise.

In other words, a restriction of ρ_δ will result the algorithm identifying all true variables as well as variables whose correlation with the true variables is higher than ρ_δ. Once we select ρ_δ, we are conceding that variables with correlations greater than ρ_δ cannot be teased apart to determine the true underlying (important) variable. By applying the correlation cut-off we are redefining our parameter. It is no longer the singular effect of A. Instead, we admit that given the data, the true important variable cannot be targeted when the data is highly

correlated and redefine our measure as a correlation-based W_δ-adjusted importance where W_δ is a newly defined subset of W based on the correlation cut-off. Given this new definition of the parameter, important variables according to the W_δ-adjusted method include all important variables as well as all variables whose correlation to a important variable is greater than a particular delta cut-off.

Therefore we must be careful when selecting ρ_δ, it must high enough to reduce bias from ETA violation, but low enough to acquire all information on the causal effects allowed by the data, which maintains the greatest level of reproducibility. If ρ_δ is higher than necessary, the list will contain decoy variables that could have been discounted using the available data. This would decrease the reproducibility of the measures in other populations. The relationship between the decoy variables and the causal variables (distribution of W) is not necessarily constant across populations while the causal mechanism (distribution of $Y|W$) can be assumed to be (i.e. the mechanisms of disease are consistent across all populations). Including decoy variables that could otherwise have been discounted adds unnecessary uncertainty when applying the final results to other populations. A method was proposed in Bembom et al. 2008 [13], which defines an analytical formula for identifying these "problem" variables data-adaptively for each A. This reduces the bias while detecting the most accurate gene set allowed by the data, maintaining reproducibility.

In this article, we apply the correlation cut-off ($\rho_\delta = \{0.5, 0.75\}$) to the breast cancer application, where the truth is unknown, and the data is noisy. In practice it is reasonable to label all potentially relevant variables as important when their effects cannot be disentangled. Setting a correlation cut-off explicitly specifies and acknowledges the method's threshold to detect the important variables among highly correlated confounders. We recommend that future applications use a larger set of ρ_δ values and provide importance measures and rankings for all variables given each ρ_δ, or data-adaptively select ρ_δ using the methods outlined in Bembom et al. 2008 [13].

6. Application

6.1. van't Veer et al (2002)

The response to standard chemotherapy among breast cancer patients can drastically vary even among women with a common stage of breast cancer at initial diagnosis. Chemotherapy is a very long and difficult treatment process, and though it is known to reduce the occurrence of metastases in 70-80% of patients, for the remaining 30-20% there is little or no response. Knowing a priori a probability of response to treatment for a given patient would aid doctors in determining a more optimal and efficient treatment plan, reducing patient discomfort and the cost of expensive trail-and-error treatment regimes. This is reflective of the current trend towards the development of individualized or "patient-tailored" treatments.

The study in [12] attempts to develop a classifier predicting treatment response to adjuvant chemotherapy among breast cancer patients based on their pre-treatment (at diagnosis) genetic profile. Given that there are over 20,000 protein-coding genes in the human genome, developing a predictor requires first reducing the data to a set of relevant genes. Here we present an application of tVIM as a method to identify these genes. Unlike linear regression and other data mining algorithms (randomForest, etc.), tVIM targets the causal effect instead of estimating only an association based on a predictive fit. We propose using tVIM to determine this subset of genes prior to the application of the prediction algorithm Super Learner [24].

The initial dataset contains 98 patients with similar stages of breast cancer at the time they enter the study. All patients are exposed to adjuvant chemotherapy. It is unknown if any other treatment methods (i.e. radiation, surgery, etc.) are applied and to what extent. For the purposes of the van't Veer analysis, the patients are assumed to be part of the same treatment arm. We continue with that assumption. Of the 98 patients, 34 develop metastases within 5 years (bad responders, Y=1), while 44 remain disease free (good responders, Y=0) [12].

6.1.1. Analysis For computation considerations we reduced our dataset to genes whose raw p-values from univariate linear regression were less than or equal to 0.05 (2254 genes) or those

which had a randomForest importance value greater than zero. We also did not include genes with more than 80% of their values missing, which left us with a total of 4446 genes. All missing data is imputed with the column mean (average gene expression over all patients). The maximum number of missing values for any gene was five.

This analysis mirrors the procedure implemented in the previous simulations. Univariate linear regression is applied to all genes. The covariate set W for each A prior to correlation cut-off includes all genes among the 4446 whose raw univariate linear regression p-value was less than or equal to 0.01 (540 genes). In application where we do not known the truth and the data is especially noisy with a complex correlation structure, we expect that we will not be able to disentangle the effects of many of the genes from one another. To minimize bias due to ETA violations, we apply a simple correlation cut-off of $\rho_\delta = \{0.5, 0.75\}$. Applying the correlation cut-off results in all potentially relevant genes labeled as important.

As in simulation we model the importance as $m(A, W_s|\beta) = \beta A$ for all A. For the initial Q we use a polynomial spline fit which allows for more complex structure of $g(W_s)$. We recommend using this or a similar data-adaptive algorithm such as DSA [25] over LARS/Lasso in application, since in reality the structure of Q may have more than just additive main effects. We also estimate $G(W_s)$ using polymars [27].

In this application the outcome is binary, therefore we interpret our tVIM measure as an approximate estimate of the excess risk.

$$m(A = a, W_s|\beta) = \mathbb{E}[Y|A = a, W_s] - \mathbb{E}[Y|A = 0, W_s] = \beta a$$

The model-based approach outlined in this paper uses standard gaussian regression for our estimate and update of $\mathbb{E}[Y|A, W_s]$. However for the purposes of variable importance, we believe the final list of ranked VIM measures and p-values are still relevant for a binary outcome. Future work is focused on the development of a more generalized model-based VIM approach which will allow us to use generalized linear regression methods.

Updated tVIM measures and p-values from t-tests are recorded and we adjust for multiple

testing using Benjamini & Hochberg (1995) step-up FDR controlling procedure [20]. In application, we recommend selecting all genes with adjusted p-values less than or equal to an appropriate cut-off (we use a standard cut-off of 0.05), and then ranking this set of genes by their absolute tVIM measure to achieve the final importance ranking of genes. Genes significant at the 0.05 level can be used as input to a prediction algorithm such as the Super Learner [24]. Results are shown in Tables I and II.

6.1.2. Results Once the univariate linear regression p-values were adjusted for multiple testing, there were no statistically significant genes at the 0.05 level, however for tVIM there were 197 and 204 genes when correlation cut-off was set at 0.5 and 0.75 respectively. In table I and table II we show the top 10 genes with the highest significance among those statistically significant at the 0.05 level.

Table I: Targeted VIM using correlation cut-off of $\rho_\delta = 0.5$: Top 10 ranked genes according to absolute importance measures among significant genes according to a p-value cut-off of 0.05.

p-value	**tVIM**	**GeneID**	**Description/Function**
0.00E+00	6.455	GALNT14 (AA165698)	UDP-N-acetyl-alpha-D-galactosamine: polypeptide N-acetylgalactosaminyltransferase 14
0.00E+00	6.164	AIP	aryl hydrocarbon receptor interacting protein
3.03E-05	3.517	LRTM1	leucine-rich repeats and transmembrane domains 1
2.33E-07	3.125	ZBTB22	zinc finger and BTB domain containing 22
6.94E-08	3.111	(AI524306)	unknown
0.00E+00	-2.843	FBXO41 (AA524093)	F-box protein 41
1.58E-06	-2.714	VAMP3	vesicle-associated membrane protein 3 (cellubrevin)
2.26E-02	-2.590	ERGIC1 (AI248720)	endoplasmic reticulum-golgi intermediate compartment (ERGIC) 1
3.27E-03	2.564	CALCOCO1	sarcoma antigen nysar3
4.38E-02	2.546	NRG2	neuregulin 2

6.1.3. Discussion Among the two top 10 lists, there are six common genes. Four of these genes, GALNT4, AIP, ZBTB22, and FBXO41, have been associated with chemotherapy

Table II: Targeted VIM using correlation cut-off of $\rho_\delta = 0.75$: Top 10 ranked genes according to absolute importance measures among significant genes according to a p-value cut-off of 0.05.

p-value	**tVIM**	**GeneID**	**Description/Function**
0.00E+00	6.455	GALNT14 (AA165698)	UDP-N-acetyl-alpha-D-galactosamine: polypeptide N-acetylgalactosaminyltransferase 14
0.00E+00	5.906	AIP	aryl hydrocarbon receptor interacting protein
0.00E+00	3.703	LRTM1	leucine-rich repeats and transmembrane domains 1
3.23E-08	3.609	(AI524306)	unknown
1.70E-08	3.331	ZBTB22	zinc finger and BTB domain containing 22
1.68E-06	-3.001	METTL1	methyltransferase like 1
6.94E-04	2.950	EIF4G1	eukaryotic translation initiation factor 4 gamma, 1
9.88E-03	2.932	SH2D3C	SH2 domain containing 3C
0.00E+00	-2.843	FBXO41 (AA524093)	F-box protein 41
1.05E-04	2.719	CTLA4	cytotoxic T-lymphocyte-associated protein 4

resistance in peer-reviewed literature. Beyond these 4, there are 3 other genes in table I with correlation cut-off at 0.5 (VAMP3, CALCOCO1, and NRG2) and 3 others in table II with correlation cut-off at 0.75 (EIF4G1, SH2D3C, and CTLA4) making it 7 out of 10 relevant genes in both lists. Variations between the two gene lists for a particular A indicate that additional genes are removed from the covariate set for when reducing the correlation cut-off from 0.75 to 0.5. After adjusting for multiple testing no genes were identified as significant based on univariate linear regression (lowest p-value 0.43).

GALNT14, which is listed first (with highest tVIM) in both tables, has been recently acknowledge as an informative biomarker for Apo2/TRAIL - based cancer therapy [29]. The Apo2/TRAIL - based cancer therapy falls into the class of apoptosis activating therapies - therapies which activate or enforce programmed cell death. Apoptosis regulates cell number in normal tissues. When apoptosis is no longer active, the tissue is considered malignant. Alternatively anthracycline, a common drug used in adjuvant chemotherapy, inhibits the topoisomerase II - alpha religation reaction leading to cytotoxic cell damage and death; while

the taxane class drugs (also common in adjuvant chemotherapy) inhibits cell division [30]. A major benefit of the Apo2/TRAIL ligand is that it preferentially induces apoptosis in cancer cells over normal cells [29]. A recent study, Wagner et al. 2007 [31], has shown that GALNT14 levels determine the sensitivity of tumor cells to apoptosis induced by Apo2L/TRAIL ligand. Increased expression of GALNT14 increases tumor cell response to this ligand making it a beneficial biomarker for sensitivity to Apo2/TRAIL - based cancer therapy. Among the patients in this study, exposed to adjuvant chemotherapy, we find GALNT14 up-regulated among the "bad-responders." Given the results of Wagner et al. 2007 [31] this could indicate that a Apo2/TRAIL - based cancer therapy may have been more beneficial for these patients. In addition to GALNT14, our results indicate that AIP, which is also known to reduce apoptosis [32, 33], is up-regulated among "bad responders" and has the second highest VIM values in both lists.

Beyond the apoptosis-related genes, we also see various indictors of drug resistance. ZBTB22 binds to Cul3 forming a complex in the Ubiquitin system and elevated Cul3 has been identified as an indicator of drug resistance [34]. The over-expression of EIF4G1 has been directly identified as an indictor of chemotherapy resistance [31]. SH2D3C interacts with BCAR and partially responsible for resistance to anti-estrogen therapy in breast cancer cells [35]. Our results indicate that all three are elevated in bad responders. In addition, CALCOCO1 has been identified as a potential target for cancer vaccines [36]. Anitbodies of CTLA-4 activate anti-tumor response in breast cancer cells. - drugs targeting this mechanism are in clinical trails [37]. NRG2 interacts with the Erbb family (including the HER-2 receptor) and induces cell growth among breast cancer cells [38]. All three again are found elevated in bad responders in our analysis. Also, FBXO41 has been found to be significant and important in numerous other biomarker discovery analyses, including ours, as an indicator of good prognosis [39]. Another interesting, though confusing result is the elevated expression of VAMP3 among "good responders." Past research has identified VAMP3 as an indicator of drug resistance [40]. It's possible that the specific chemotherapy treatment chosen was correct for patients with elevated VAMP3. Specifics are unknown.

7. Conclusion

In both simulation and application we see the necessity for a standard method. Results vary widely leading to long lists and confusion, which list to use? In this paper we propose using tVIM as a standard method for biomarker discovery. In simulation it has proven resilient to increases in correlation, controlling type I error. It also provides an interpretable and meaningful measure of importance, which given an appropriate study design is interpretable as an estimate of a causal effect.

By targeting the causal effect, the measures obtained by tVIM are less sensitive to changes in the covariate distribution and therefore more reproducible in any population given it has the same conditional distribution of $Y|W$. This allows tVIM measures to be generalizable across microarray platforms that may have different noise levels. This reproducibility is essential for any standardized method, increasing confidence in diagnostic and treatment decisions based on these measures. In other words, if the causal effect between gene A and the response is correctly estimated in a population, it will be applicable to other populations. If instead we attribute the effect to gene B which is highly correlated to the causal gene A in the first population the correlation between gene B and gene A is not necessarily consistent in the other populations making the measure effect inapplicable in those populations. For instance if people in the second population have a cold, and gene B is related to immune response. It's levels may be much higher and no longer correlated in the same degree with the level of gene A. Making inferences on the disease state from the level of gene B erroneous.

In comparison, common univariate linear regression is highly susceptible to increases in type I error due to increased correlation among variables. And though LASSO/LARS provides improvement, using tMLE to update its estimate increases the accuracy in the importance estimate and rank (See Appendix (B)) and provides the correct asymptotic inference [9].

In application, tVIM identifies genes biologically related to chemotherapy resistance as well as genes which indicate a possible mechanism of treatment for "poor responders" based on up-to-date biological information. These promising results and the relevance of the gene list

supports the use of tVIM for biomarker discovery and as a pre-screening method for prediction. It is simple to implement and understand and is adaptable most data types including binary variables, survival outcome, and longitudinal data [6]. This accuracy, reproducibility, and flexibility of the tVIM method make it an strong candidate for a standardized biomarker discovery method.

Future work for tVIM will focus on developing variable importance methods for non-gaussian outcomes (binary Y), as well as methods for identifying the best correlation cut-off for each variable, A. Applying a correlation cut-off in practice reduces the bias in the tVIM estimate due to potential ETA violations. However, the difference between the lists for tVIM correlation cut-off 0.5 and 0.75 affirm the need for a method which identifies the proper cut-off for a given gene. Having too low of a cut-off neglects controlling for the appropriate genes to achieve an estimate of the causal effect, decreasing its reproducibility across populations. Having too high a cut-off leads to ETA violations which increase bias in our importance estimate. We also will explore methods which help piece apart or at the very least elucidate the relationship among a group of heavily correlated variables in relation to a response.

ACKNOWLEDGEMENTS

This work was done under the grant for Targeted Empirical Super Learning in HIV Research, funding through NIH National Institute of Allergy and Infectious Diseases; Award number R01 A1074345-01

APPENDIX

I. Targeted Maximum Likelihood

Targeted Maximum Likelihood (tMLE) methodology maximizes the likelihood in a direction which targets the parameter of interest using the appropriate bias-variance trade-off tMLE.

We defined

$$\mu(a) = \mathbb{E}_W[m(A = a, W|\beta)]$$

with the estimate at a particular A=a defined as

$$\mu(a) = \frac{1}{n}\sum_{i=1}^{n}[m(a, W_i|\beta)]$$

where m(.) models the effect

$$m(a, W|\beta) = \mathbb{E}_P[Y|A = a, W] - \mathbb{E}_P[Y|A = 0, W])$$

When A is binary, IPTW and DR-IPTW [9, 6] methods may be used to estimate $\mu(A)$ without model assumptions. When A is more general, it requires specification of a model $m(A, W|\beta(P))$ that satisfies $m(A = 0, W|\beta(P)) = 0$, where the true $\beta_0 = \beta(P_0)$.

Targeted MLE methodology creates a path through the true density p^0, represented as the hardest sub-model $p^0(\epsilon)$. The hardest submodel $p^0(\epsilon)$ is selected to only vary $Q(p)(Y|A, W)$, with score equal to $D_h(p_n^0)$ at $\epsilon = 0$. This sub-model is explicitly derived in [6].

Where $D_h(p_n^0)$ is the efficient influence curve as defined according to the following theorem

Theorem 1. *(From Yu et al. 2003 [16]) For parameter $p \to \beta(p)$ in model $M = \{p : \mathbb{E}_p(Y|A, W) - \mathbb{E}_p(Y|A = 0, W) = m(0, W|\beta(p))$, satisfying $m(0, W|\beta) = 0$ for all $\beta \in I\!R^d$ the orthogonal complement of the nuisance tangent space is*

$$T_{nuis}^{\perp}(p) = \{D_h(p) : h\}$$

where

$$D_h(p)(O) \equiv \{h(A, W) - \mathbb{E}_p(h(A, W)|W)\}(Y - m(A, W|\beta(p)) - \mathbb{E}_p(Y|A = 0, W))$$

The efficient influence curve or canonical gradient is then defined as

$$D_{h_{opt}}(p)(O) = \{h_{opt}(A, W) - \mathbb{E}_p(h_{opt}(A, W)|W)\}(Y - m(A, W|\beta(p)) - \mathbb{E}_p(Y|A = 0, W))$$

where

$$h_{opt} = \frac{1}{\sigma(A, W)}\left\{\frac{d}{d\beta}m(A, W|\beta) - \frac{\mathbb{E}\left[\frac{1}{\sigma(A,W)}\frac{d}{d\beta}m(A, W|\beta)|W\right]}{\mathbb{E}\left[\frac{1}{\sigma(A,W)}|W\right]}\right\}$$

and $Var(Y|AW) = \sigma(A, W)$. If we assume $Var(Y|A, W) = Var(Y|W)$, then a more practical form

of h_{opt} is available

$$h^*_{opt} = \frac{1}{\sigma(A,W)} \left\{ \frac{d}{d\beta} m(A,W|\beta) - \mathbb{E}\left[\frac{d}{d\beta} m(A,W|\beta) | W \right] \right\}$$

where $G(W) = \mathbb{E}(A = a|W)$ and $Q(a,W) = \mathbb{E}_P(Y|A = a, W)$ are nuisance parameters. The double robust nature of the estimating function gives $\mathbb{E}_{P_0} D(O|\beta_0, Q, G) = 0$, providing a consistent estimate of β, if either of the nuisance parameters ($G(W)$ and $Q(A,W)$) is specified correctly.

Note when $\rho > 0$, the experimental treatment assumption (ETA) (i.e. $P(W_j|W_{-j}) = P(W_j)$) no longer holds. However due to the nature of the simulated data where all variables are simulated from a multivariate normal, the dependency can be accurately modeled using a main term linear model due to the simple correlation structure (i.e. $\mathbb{E}(W_j|W_{-j}) = \beta_W W_{-j}$).

Given an initial estimate of the density $p^0_n = Q^0(A,W)$, and defining the hardest sub-model $p^0(\epsilon|p^0_n)$, $p^0(\epsilon|p^0_n)$ is maximized with respect to ϵ, substituting in the new estimate ϵ_n, the updated density $p^1 = p^0(\epsilon_n|p^0_n)$, is the new targeted density. In some cases iteration is necessary (substituting the new density estimate as initial density estimate and solving again for ϵ). By maximizing $p^0(\epsilon|p^0_n)$ for ϵ, tMLE maximizes the likelihood in the direction of the parameter of interest μ, making the final density estimate the solution to $P_n(D_h(O)) = 0$ as well.

Assuming a normal distribution for $Q(p^0_n)(Y|A,W)$ with mean $Q(p^0_n)(A,W) = \mathbb{E}_{p^0_n}(Y|A,W)$ and variance $\sigma^2(Q^0_n)(A,W)$, the hardest sub-model which updates the original $Q(p^0_n)(Y|A,W)$ in a direction which estimates the parameter of interest well, can be defined as

$$Q(p)(\epsilon)(Y|A,W) = f_0\left(\frac{Y - m(A,W|\beta^0_n(\epsilon)) - Q^0_n(\epsilon)(W)}{\sigma(A,W)} \right)$$

where f_0 represents the standard normal density with updated parameters $\beta^0_n(\epsilon) = \beta^0_n(Q^0_n) + \epsilon$ and $\theta^0_n(\epsilon) = \theta^0_n(Q^0_n) + \epsilon^T r(W)$ where

$$r(p^0_n)(W) = \frac{\mathbb{E}\left[\frac{1}{\sigma(A,W)} \frac{d}{d\beta} m(A,W|\beta) | W \right]}{\mathbb{E}\left[\frac{1}{\sigma(A,W)} | W \right]}$$

if we assume, with only some loss in efficiency that $\sigma(A,W) = \sigma(W)$, then the above reduces to

$$r^*(p^0_n)(W) = \mathbb{E}\left[\frac{d}{d\beta} m(A,W|\beta) | W \right]$$

The proper form of r(W) shown above is found by equating the score of $Q(p)(\epsilon)$ in terms of ϵ at

$\epsilon = 0$ to the efficient influence curve $D_{h_{opt}}(p_n^0)$.

The likelihood for $Q(p)(\epsilon)$ can now be maximized for ϵ using standard weighted least squares, updating the initial estimates of β and θ, providing the new targeted estimate of the overall density as well as the parameter of interest β. Given a linear model $m(A, W|\beta)$ in β, a closed form solution for ϵ does exist and standard weighted linear regression software packages such as lm() in R may be used.

I.1. Inference and Testing

Asymptotically tMLE is equivalent to solving

$$\mathbb{E}_P(D_h(O|\beta, Q, G)) = 0$$

where $D_h(O|\beta, Q, G)$ is the efficient influence curve, making formal inference dependent on the influence curve still applicable to TMLE derived estimates.

The covariance matrix for β can be estimated using the conservative influence curve, The conservative influence curve is defined as,

$$IC(O) = \frac{D(O|\beta_0, \Pi, \theta)}{\mathbb{E}[\frac{d}{d\beta} D(O|\beta_0, \Pi, \theta)]}$$

where

$$\sqrt{n}(\beta_n - \beta_0) \sim N(0, \Sigma_n)$$

asymptotically with covariance equal to

$$\Sigma_n = \frac{1}{n} \sum \hat{IC}(O) \hat{IC}(O)^T$$

Covariance can also be estimated by bootstrap estimates of β, but this would requiring extra computational time. In this study, we know that $G(W)$ is correct, therefore estimates based on the influence curve are consistent.

Testing $H_0 : \beta_0(j) = 0$, p-values can be determined using test statistic

$$T_n(j) = \frac{\sqrt{n}\beta_n(j)}{\sqrt{\Sigma_n(j,j)}} \underset{n\to\infty}{\sim} N(0, 1)$$

Testing for significance of the marginal variable importance curve when the effect of A is modified

by W_i (i.e $\mathbb{E}[m(A = a, W = a|\beta)] = \beta_0 a + \beta a \mathbb{E}[W_i]$) is completed by testing the null hypothesis $H_0 : c^T \beta_0 = 0$, where c is the appropriate vector of A and W corresponding to m(.). Test statistic becomes $T_n(j) = \frac{\sqrt{n} c^T \beta_n(j)}{\sqrt{c^T \Sigma_n(j,j) c}}$ which is asymptotically distributed N(0,1).

II. Additional Simulation results

Previous performance measures are focused on determining how well the methods rank the true variables with respect to all variables. Average importance measures showcase the ability of each method to not only distinguish true variables from decoys, but also properly determine the magnitude of importance accurately

The average importance value is plotted versus actual value for LM, Q (LASSO), and tVIM methods at each correlation level. This is only relevant for LM, Q, and tVIM, which are on the same scale as the simulated importance measures.

When the actual importance values are easily distinguished for the 10 truly dependent variables, such as when $\beta = 1, ..., 10$, we can distinguish the importance of the true variables relative to other true variables. When $\beta = 1, ..., 10$, average rank and importance should lie on the x=y line when plotting average rank or measure by true importance value.

The difference between the true measure/rank versus the estimated average measure/rank is summarized by calculating the mean squared deviation of the estimated values from the true values. These measures are plotted versus correlation providing a visual representation of the effect correlation among the covariates has on the overall accuracy of each method.

II.1. Average Importance Value

II.1.1. Average Importance Rank We can see clearly that the variables with $\beta = 1$ and 2 are harder for all methods to pinpoint and rank accurately.

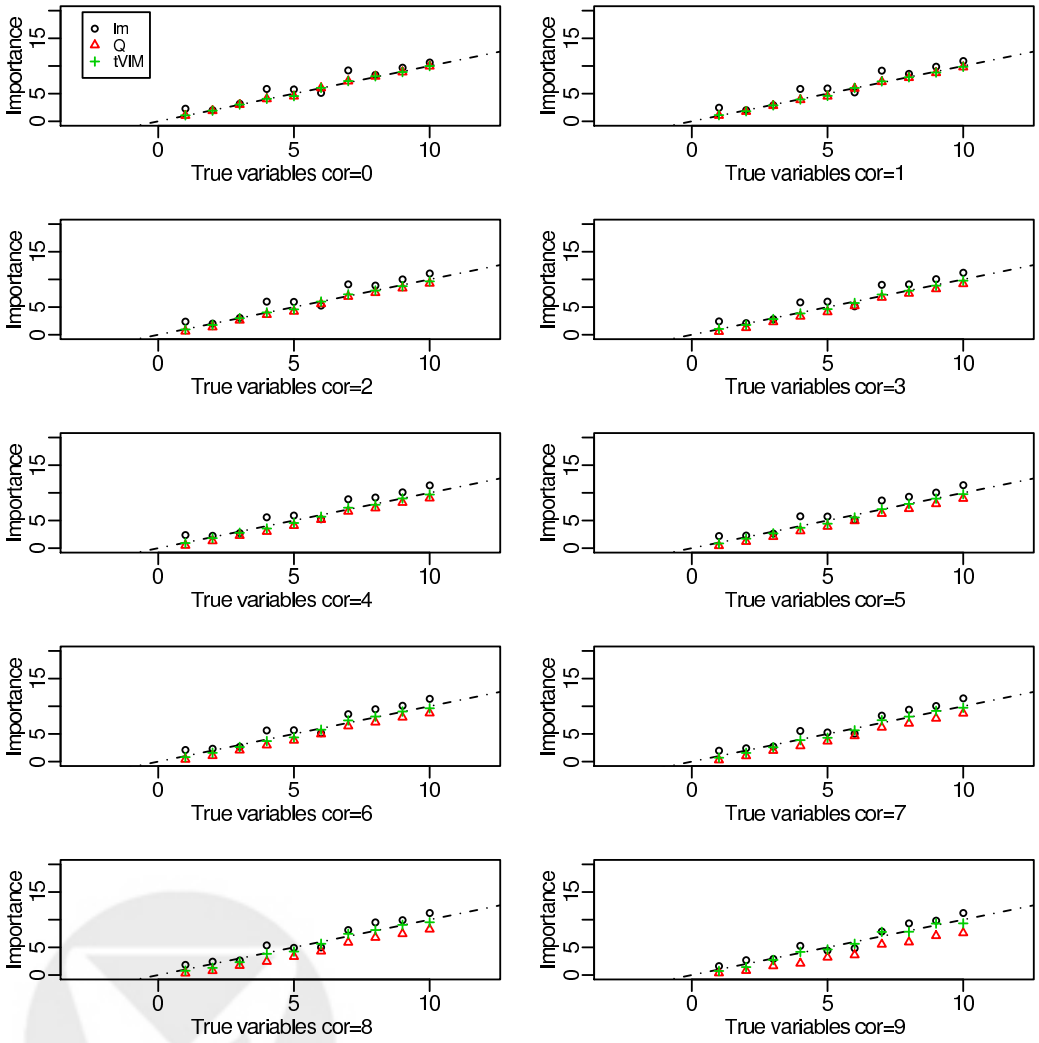


Figure 6: Average importance value for each of ten true variables with importance values = 1,...,10. Plots included for all $\rho = 0, ..., .9$. Only linear regression, LASSO, and tVIM are analyzed since RF values are not necessarily on the same scale as the true level of importance. ($\sigma_Y = 10$)

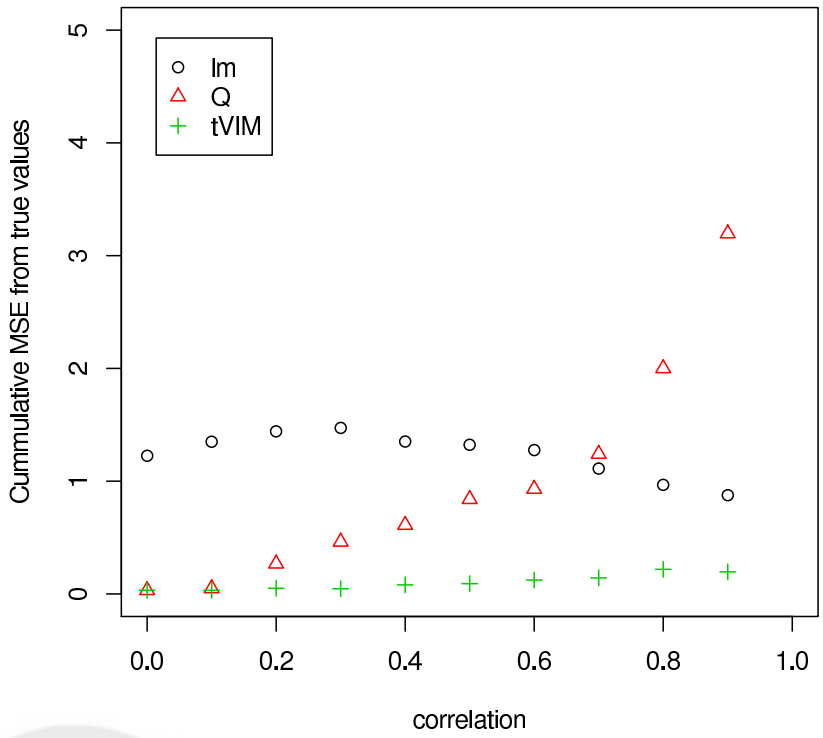


Figure 7: Mean square error difference between average importance values and true values at $\rho = 0, ..., .9$. This relates to Figure 7.

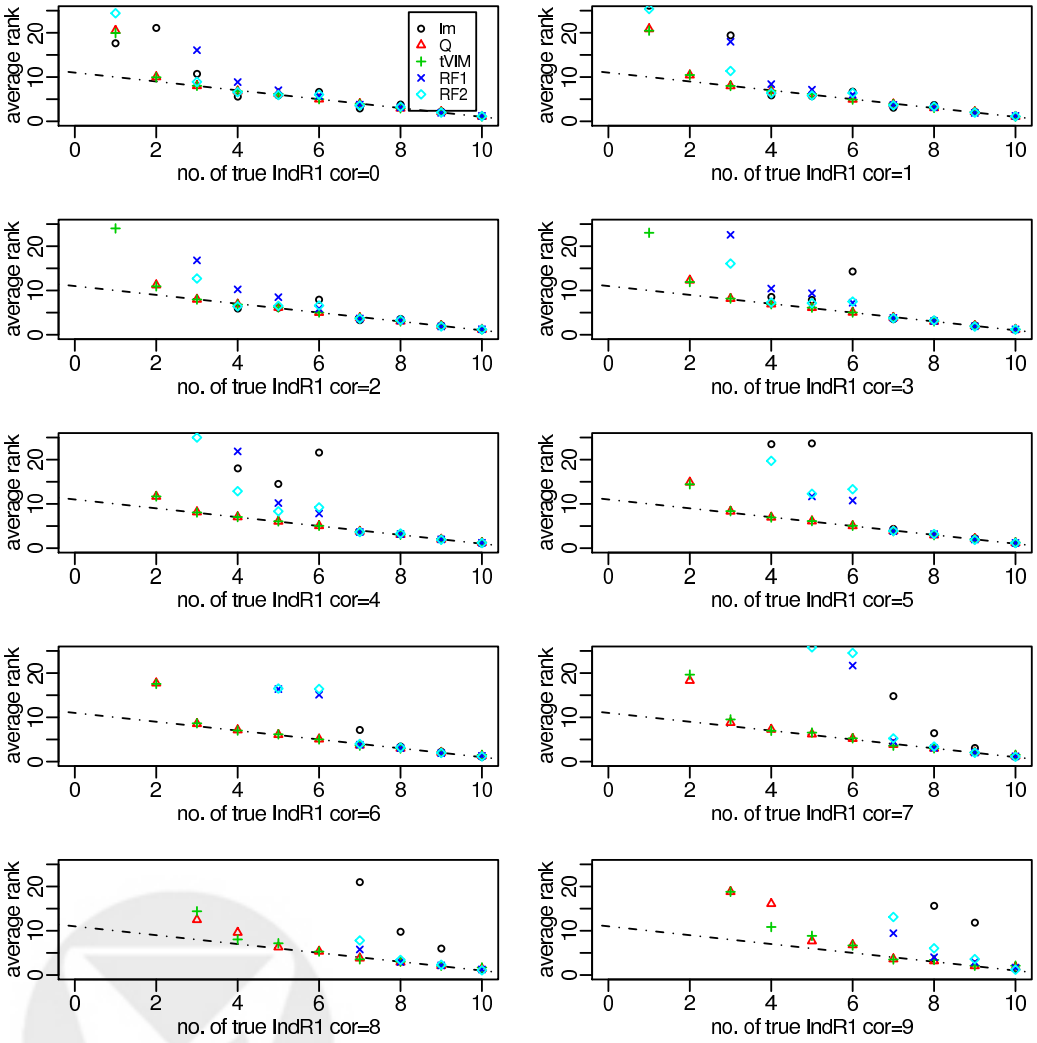


Figure 8: Average importance rank for each of ten true variables with actual ranks = 1,...,10. Plots included for all $\rho = 0, ..., .9$, ranking by measure. ($\sigma_Y = 10$). Comparing the methods by rank allows us to include RF1 and RF2 in the comparison even though their measures are on a different scale

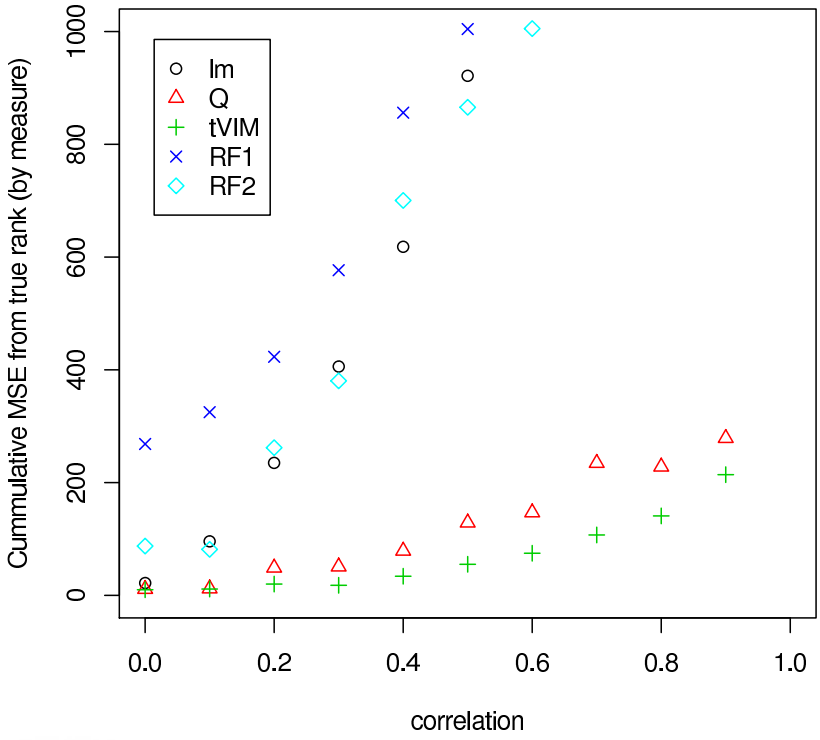


Figure 9: Mean square error difference between average importance ranks and true ranks at $\rho = 0, ..., .9$, when ranking by measure. This relates to Figure 9. Comparing the methods by rank allows us to include RF1 and RF2 in the comparison even though their measures are on a different scale.

REFERENCES

1. U.S. Department of Health and Human Services and U.S. Food and Drug Administration. Critical Path Initiative Fact Sheet. *http://www.fda.gov/oc/initiatives/criticalpath/factsheet.html* 2007.
2. U.S. Department of Health and Human Services and U.S. Food and Drug Administration. Critical Path Opportunities Report. *http://www.fda.gov/oc/initiatives/criticalpath/reports/opp$_r$eport.pdf* 2006.
3. Pearl J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.
4. Wang Y, Petersen ML, Bangsberg D, van der Laan MJ. Diagnosing bias in the inverse probability of treatment weighted estimator resulting from violation of experimental treatment assignment. Tech. Rep. Working Paper 211, U.C. Berkeley Division of Biostatistics Working Paper Series., September 2006. http://www.bepress.com/ucbbiostat/paper211.
5. Breiman L. Random forests. *Machine Learning* 2001; **45**(1):5–32. http://dx.doi.org/10.1023/A:1010933404324.
6. van der Laan MJ, Rubin D. Targeted maximum likelihood learning. Working paper 213, U.C. Berkeley Division of Biostatistics Working Paper Series, October 2006. http://www.bepress.com/ucbbiostat/paper213.
7. Pollard K.S, van der Laan M.J. A New Algorithm for Hierarchical Hybrid Clustering with Visualization and the Bootstrap. *Journal of Statistical Planning and Inference* 2003; **117**:275303.
8. Bembom O, Petersen ML, Ree SY, Fessel WJ, Sinisi SE, Shafer RW, van der Laan MJ. Biomarker discovery using targeted maximum likelihood estimation: Application to the treatment of antiretroviral resistant HIV infection. *Statistics in Medicine* 2009; **28**(1):152–172.
9. van der Laan MJ. Statistical inference for variable importance. Tech. Rep. Working Paper 188, U.C. Berkeley Division of Biostatistics Working Paper Series, 2005. http://www.bepress.com/ucbbiostat/paper188.
10. Efron B, Hastie T. lars. R package.
11. Liaw A, Wiener M. randomforest. R package.
12. van 't Veer L, Dai H, van de Vijver M, He Y, Hart A, Mao M, Peterse H, van der Kooy K, Marton M, Witteveen A, Schreiber G, Kerkhoven R, Roberts C, Linsley P, Bernards R, Friend S. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002; **415**(6871):530–6.
13. Bembom O, Fessel JW, Shafer RW, van der Laan MJ. Data-adaptive selection of the adjustment set in variable importance estimation. Tech. Rep. Working Paper 231, U.C. Berkeley Division of Biostatistics Working Paper Series, March 2008. http://www.bepress.com/ucbbiostat/paper231.
14. Robins J, Mark S, Newey W. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* 1992; **48**(479-495).
15. Robins J, Rotnitzky A. Comment on the bickel and kwon article "inference for semiparametric models: Some questions and an answer". *Statistica Sinica* 2001; **11**(4):920–936.
16. Yu Z, van der Laan MJ. Measuring treatment effects using semiparametric models. Tech. Rep. Working Paper 136, U.C. Berkeley Division of Biostatistics Working Paper Series, September 2003.

http://www.bepress.com/ucbbiostat/paper136.

17. Tuglus C, van der Laan MJ. Targeted methods for biomarker discovery: The search for a standard. Tech. Rep. Working Paper 233, U.C. Berkeley Division of Biostatistics Working Paper Series, March 2008. http://www.bepress.com/ucbbiostat/paper233.
18. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Annals of Statistics (with discussion)* 2004; **32**(2):407–499.
19. Tibshirani R. Regression shrinkage and selection via the lasso. *J. Royal Statist. Soc B.* 1996; **58**(1):267–288.
20. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B.* 1995; **57**:289–300.
21. Pollard K, Dudoit S, van der Laan M. *Multiple Testing Procedures: R multtest Package and Applications to Genomics in Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. 209-229. Springer (Statistics for Biology and Health Series), 2005.
22. Flach PA. Tutorial on "the many faces of roc analysis in machine learning". In *The Twenty-First International Conference on Machine Learning*.
23. Carey V. Roc. R package.
24. van der Laan MJ, Polley EC, Hubbard AE. "super learner". Tech. Rep. Working Paper 222, U.C. Berkeley Division of Biostatistics Working Paper Series, July 2007. http://www.bepress.com/ucbbiostat/paper222.
25. Sinisi SE, van der Laan MJ. Loss-based cross-validated deletion/substitution/addition algorithms in estimation. Working paper 143, U.C. Berkeley Division of Biostatistics Working Paper Series, March 2004. http://www.bepress.com/ucbbiostat/paper143.
26. van der Laan MJ, Robins JM. *Unified Methods for Censored Longitudinal Data and Causality*. Springer, 2003.
27. O'Connor M. polymars. R package polspline.
28. Kooperberg C, Bose S, Ston CJ. Polychotomous regression. *Journal of the American Statistical Association* 1997; **92**:117–127.
29. Fesik SW. Promoting apoptosis as a strategy for cancer drug discovery. *Nature Reviews Cancer* 2005; **5**:876–885.
30. Cortes-Funes H, Coronado C. Role of anthracyclines in the era of targeted therapy. *Cardiovasc Toxicol* 2007; **7**:56–60.
31. Wagner KW, Punnoose EA, Januario T, Lawrence DA, Pitti RM, Lancaster K, Lee D, von Goetz M, Yee SF, Totpal K, Huw L, Katta V, Cavet G, Hymowitz SG, Amler L, Ashkenazi A. Death-receptor o-glycosylation controls tumor-cell sensitivity to the proapoptotic ligand apo2l/trail. *Nat Med* 2007; **13**(9):1070–1077. http://dx.doi.org/10.1038/nm1627.
32. Vogel C, Li W, Sciullo E, Newman J, Hammock B, Reader J, Tuscano J, Matsumura F. Pathogenesis of aryl hydrocarbon receptor-mediated development of lymphoma is associated with increased cyclooxygenase-2 expression. *Am J Pathol.* 2007; **171**(5):1538–48.

33. Berwick M, Matullo G, Song YS, Guarrera S, Dominguez G, Orlow I, Walker M, Vineis P. Association between aryl hydrocarbon receptor genotype and survival in soft tissue sarcoma. *Journal of Clinical Oncology* 2004; **22**(19):3997–4001.

34. Zhang H, Tomida A, Koshimizu R, Ogiso Y, Lei S, Tsuruo T. Cullin 3 promotes proteasomal degradation of the topoisomerase i-dna covalent complex. *Cancer Res.* 2004; **64**(3):1114–21.

35. Near RI, Zhang Y, Makkinje A, Borre PV, Lerner A. And-34/bcar3 differs from other nsp homologs in induction of anti-estrogen resistance, cyclin d1 promoter activation and altered breast cancer cell morphology. *J Cell Physiol.* 2007; **212**(3):655–65.

36. Lee S, Obata Y, Yoshida M, Stockert E, Williamson B, Jungbluth A, Chen Y, Old L, Scanlan M. Immunomic analysis of human sarcoma. *Proc Natl Acad Sci U S A.* 2003; **100**(5):2651–6.

37. Korman A. Ctla-4 based therapy (mdx-010). *Breast Cancer Res* 2003; **5**(Suppl 1):63.

38. SweeneyDagger C, Fambrough D, Huard C, Diamont AJ, Lander ES, CantleyDagger LC, Carraway KL. Growth Factor-specific Signaling Pathway Stimulation and Gene Expression Mediated by ErbB Receptors. *J. Biol. Chem.* 2001; **276**(25):22685–22698.

39. Alexe G, Alexe S, Axelrod DE, Bonates TO, Lozina II, Reiss M, Hammer PL. Breast cancer prognosis by combinatorial analysis of gene expression data. *Breast Cancer Research* 2006; **8**(4):R41.

40. Nakamura Y, Katagiri T, Nakatsuru S. (wo/2005/028676) method of diagnosing breast cancer. Patent, Oncotherapy Science, Inc., 2005. http://www.wipo.int/pctdb/en/wo.jsp?ELEMENT$_SET$ = $FLANGUAGE = ENGKEY = 05$

7.2 *Biomarker Discovery using Targeted Maximum Likelihood Estimation: Application to the Treatment of Antiretroviral Resistant HIV Infection*

The following article appears as it was originally published on the University of California, Berkeley Division of Biostatistics Working Paper Series website in 2007, http://www.bepress.com/ucbbiostat/paper221/.

It was later published in *Statistics in Medicine* in 2008, http://www3.interscience.wiley.com/journal/121422393/abstract.

Biomarker discovery using targeted maximum likelihood estimation: Application to the treatment of antiretroviral resistant HIV infection.

Oliver Bembom[1], Maya L. Petersen[1], Soo-Yon Rhee[2], W. Jeffrey Fessel[3], Sandra E. Sinisi[1], Robert W. Shafer[2], and Mark J. van der Laan[1]

[1]Division of Biostatistics, University of California, Berkeley, CA
[2]Division of Infectious Diseases,Center for AIDS Research, Stanford University, Palo Alto, CA
[3]Clinical Trials Unit, Kaiser Permanente, San Francisco, CA

Abstract

Researchers in clinical science and bioinformatics frequently aim to learn which of a set of candidate biomarkers is important in determining a given outcome, and to rank the contributions of the candidates accordingly. This article introduces a new approach to research questions of this type, based on targeted maximum likelihood estimation of variable importance measures.

The methodology is illustrated using an example drawn from the treatment of HIV infection. Specifically, given a list of candidate mutations in the protease enzyme of HIV, we aim to discover mutations that reduce clinical virologic response to antiretroviral regimens containing the protease inhibitor lopinavir. In the context of this data example, the article reviews the motivation for covariate adjustment in the biomarker discovery process. A standard maximum likelihood approach to this adjustment is compared with the targeted approach introduced here. Implementation of targeted maximum likelihood estimation in the context of biomarker discovery is discussed, and the advantages of this approach are highlighted. Results of applying targeted maximum likelihood estimation to identify lopinavir resistance mutations are presented and compared with results based on unadjusted mutation-outcome associations as well as results of a standard maximum likelihood approach to adjustment.

The subset of mutations identified by targeted maximum likelihood as significant contributors to lopinavir resistance is found to be in better agreement with current understanding of HIV antiretroviral resistance than the corresponding subsets identified by the other two approaches. This finding suggests that targeted estimation of variable importance represents a promising approach to biomarker discovery.

1 Introduction

Researchers in bioinformatics, biostatistics, and related fields are often faced with a large number of candidate biomarkers and aim to assess their importance in relation to a given outcome. Examples include the identification of single nucleotide polymorphisms associated with the development of cancers, identification of HLA types associated with disease progression rates, and the identification of viral mutations that contribute to reduced susceptibility to drug therapy. In some cases, the goal may be to select from a list of candidates those biomarkers with underlying causal relationships to the outcome. In others, the researcher may wish to rank the importance of a set of candidate biomarkers in terms of their contributions to determining the outcome.

In this article we introduce a novel method for biomarker discovery based on targeted maximum likelihood estimation of variable importance measures (VIMs) [15]. As we discuss, the marginal association of a candidate biomarker with the outcome may not reflect the biomarker's mechanistic or prognostic significance. For example, a viral mutation may be associated with poor response to a given drug without playing any mechanistic role in resistance, as a result of covariates that both predict the presence of the mutation and affect the outcome via an alternative pathway. VIMs provide a means to rank candidate biomarkers based on their association with a given outcome, controlling for a large number of additional covariates [13]. Specifically, given a binary candidate biomarker A, an outcome Y, and a list of covariates W, the W-adjusted VIM is defined as $E_W(E(Y|A=1,W) - E(Y|A=0,W))$. If one is willing to assume that the measured covariates W are sufficient to control all confounding of the effect of A on Y, then the VIM can be interpreted as the average causal effect of the biomarker on the outcome. In the absence of such an assumption, the VIM remains an interpretable summary measure of the importance of the biomarker after controlling for specified covariates.

Several approaches are available to estimate VIMs. Perhaps the most common approach is based on maximum likelihood estimation of the conditional expectation of the outcome given the candidate biomarker and covariates. This conditional expectation is then evaluated at $A = 1$ and $A = 0$ for each subject, and the difference is averaged across the population. Such an approach corresponds to the G-computation formula of Robins [9] applied at a single time point.

In this article, we show how a recent advance in statistical methodology, targeted maximum likelihood estimation, can improve on this standard approach. Targeted maximum likelihood estimation involves a simple one-step adjustment to an initial estimate of the conditional expectation of the outcome given the biomarker and covariates. This adjustment reduces bias in the estimate of the VIM and improves robustness to mis-specification of the likelihood. The theoretical basis for targeted maximum likelihood estimation was recently published by van der Laan and Rubin [15]. Here, we demonstrate how this work can be applied in practice to improve standard approaches to biomarker discovery. Throughout the article, emphasis is placed on practical understanding and implementation of the methods described.

Targeted maximum likelihood is illustrated using an original data example drawn from the treatment of antiretroviral resistant HIV-infection. Using observational clin-

ical data, we aimed to determine which of a set of candidate viral mutations affect clinical virologic response to the antiretroviral drug lopinavir, and to rank the importance of these mutations for drug-specific resistance. The resulting ranking can be used to inform interpretation of viral genotypes, and to aid clinicians in selecting new antiretroviral treatment regimens with a greater probability of virologic success.

1.1 Outline.

The article has the following structure. Section 2 introduces the data application and provides background on the research question and the data structure. In Section 3, we discuss methods for biomarker discovery, and compare estimation of unadjusted and adjusted associations between the candidate biomarker and the outcome ($E(Y|A = 1) - E(Y|A = 0)$ and $E_W(E(Y|A = 1, W) - E(Y|A = 0, W))$, respectively). Section 4 presents the targeted maximum likelihood approach to estimation of W-adjusted VIMs, and compares it to a standard (or G-computation) approach. Implementation and inference using the targeted approach are discussed both generally and in the context of the data example. Section 5 presents the results of the data analysis, in which the importance of candidate mutations was assessed using unadjusted, G-computation, and targeted estimates of VIMs. We compare the results of these methods, and discuss them in the context of current understanding of HIV antiretroviral resistance. Section 6 concludes with a discussion.

2 Application: Identification of HIV mutations associated with decreased viral susceptibility to lopinavir.

2.1 Research Question.

Virus resistant to antiretroviral drugs frequently evolves during treatment of HIV infection and can result in disease progression if new therapies are not initiated. Designing an effective salvage therapy regimen for an individual infected with resistant virus requires choosing drugs to which the virus infecting that individual remains sensitive. Tests of viral resistance are now available to help guide salvage regimen design. However, interpretation of the results of resistance tests for the purposes of guiding salvage regimen drug choice remains complex.

Assays of viral susceptibility to antiretroviral drugs fall into two general categories: phenotype-based and genotype-based. Phenotypic resistance tests directly quantify *in vitro* drug susceptibility using recombinant virus, while genotypic resistance tests are performed by sequencing the genes for the viral protease and reverse transcriptase enzymes, the targets of the major antiretroviral classes. While genotypic tests are less expensive, less complex, and faster to perform than phenotypic tests, interpretation of the results of genotypic tests requires linking patterns of viral mutations to *in vivo* and *in vitro* resistance.

Data from several sources have been used to inform interpretation of viral genotype. Observed associations between the presence of specific viral mutations and patients' treatment histories suggest that these mutations have been selected for over the course of therapy and likely contribute resistance to the specific drugs used. *In vitro* experiments have also provided insight into the role of individual mutations in determining drug-specific viral susceptibility. Such experiments include observation of viral evolution in the presence of antiretroviral drugs, and tests of the ability of mutated viruses to replicate in the presence of drug. The resulting data on links between viral mutations and susceptibility to antiretroviral drugs have been combined to create rule-based algorithms for the interpretation of genotype data. Examples include the French ANRS (National Agency for AIDS Research) algorithm [4], the Rega algorithm [7], and the Stanford HIVdb program [11]. The Stanford algorithm in particular provides drug-specific estimates of viral susceptibility using a weighted scoring system for mutations thought to be associated with resistance. Viral susceptibility to an entire regimen is calculated by summing susceptibility scores for each drug in the regimen, yielding a genotypic susceptibility score (GSS). The International AIDS society (IAS) also publishes an annual drug-specific list of mutations thought to affect viral resistance [6].

Ultimately, the goal of such algorithms is to identify mutations with large impacts on clinical drug response. We aimed to use data from an observational clinical cohort to rank a list of candidate resistance mutations based on their importance in conferring resistance to specific antiretroviral drugs. For the sake of illustration, we focused on resistance to the commonly used protease inhibitor (PI) drug lopinavir. Rankings like the one presented here can be used to inform current genotype interpretation algorithms, with the aim of improving selection of salvage antiretroviral drug regimens for patients infected with resistant HIV virus.

2.2 Data.

Study sample and inclusion criteria.

Analyses were based on observational clinical data that were primarily drawn from the Stanford drug resistance database and supplemented with data from an ongoing collaboration with the Kaiser Permanente Medical Care Program, Northern California. Currently, the Stanford database contains longitudinal data on over 6,000 patients. Data collected include use of antiretroviral drugs, results of viral genotype tests, and measurements of plasma HIV RNA level (viral load) and CD4 T cell count collected during the course of clinical care.

We identified all Treatment Change Episodes (TCEs) in this database that involved initiation of a salvage regimen containing lopinavir. A TCE was defined using the following inclusion criteria: 1) change of at least one drug from the patient's previous antiretroviral regimen; 2) availability of a baseline viral load and genotype within 24 weeks prior to the change in regimen; and, 3) availability of an outcome viral load 4-36 weeks after the change in regimen and prior to any subsequent changes in regimen.

TCEs were excluded if no candidate resistance mutations were present in the baseline genotype, if the subject had no past experience of PI drugs prior to the current regimen, or if the newly initiated regimen included hydroxyurea, any experimental an-

tiretroviral drugs, or any PI drugs other than lopinavir (apart from the low dose of ritonavir that is always given with lopinavir). If a single baseline genotype had several subsequent regimen changes that met inclusion criteria as TCEs, only the first of these regimen changes was included in analyses. Multiple TCEs, each corresponding to a unique baseline genotype, treatment changes, and outcome, were allowed from a single individual; the resulting dependence between TCEs was accounted for in the derivation of standard errors and p-values.

Data structure.

Baseline genotype was summarized as a vector $\mathbf{A}$ of binary variables A_j that indicate the presence of a specific mutation in the protease enzyme of HIV (the viral target of lopinavir). We considered as candidate biomarkers all mutations assessed by the Stanford HIVdb algorithm to be potentially related to resistance to any approved PI drug (`http://hivdb.stanford.edu`, accessed 7/18/2006). In total, we considered 30 candidate PI mutations. In the sections that follow, we describe methods for estimating the importance of a single candidate biomarker A. In applying these methods to the data example, each of the candidate mutation A_j, for $j = 1, ..., 30$, was assessed separately; however, for simplicity we suppress the subscript j.

Antiretroviral regimens generally combine drugs from more than one class. The following characteristics of the non-PI component of the salvage regimen were included in the set W of adjustment variables: indicators of use of each of 13 non-PI drugs; number of drugs used in each major non-PI class (nucleoside reverse transcriptase inhibitors or NRTI, and non-nucleoside reverse transcriptase inhibitors or NNRTI); number of drugs and number of classes used in the salvage regimen for the first time; use of an NNRTI drug in the salvage regimen for the first time; and number of drugs switched between the previous and salvage regimen.

W also included the following covariates collected prior to the baseline genotype: indicators of past treatment with each of 30 antiretroviral drugs; number of drugs used in each of the three major drug classes (PI, NRTI, and NNRTI); history of mono or dual therapy; number of past drug regimens; date of earliest antiretroviral therapy; highest prior viral load; lowest prior CD4 T cell count; and most recent (baseline) viral load.

Summaries of non-PI mutations in the baseline genotype (i.e. mutations in the reverse transcriptase enzyme targeted by the NRTI and NNRTI classes) were also included in the covariate set W. Known NRTI and NNRTI resistance mutations present at baseline were summed. In addition, susceptibility scores (standardized to a 0-1 scale) were calculated for each non-PI antiretroviral drug using the Stanford HIVdb scoring system. These susceptibility scores were included both as individual covariates and as interactions with indicators of the use of their corresponding drugs in the salvage regimen. Finally, these interaction terms were summed to yield a non-PI GSS, which summarized the activity of the non-PI component of the regimen.

The outcome of interest, clinical virologic response, could be conceived as either a binary indicator of success (defined as achievement of a final viral load below the assay's lower limit of detection of 50 copies/mL), or as a continuous measure such as the change in final log_{10} viral load over baseline log_{10} viral load. The analyses reported here used

a hybrid of these two approaches, aiming to capture the strengths of each. Specifically, given a baseline measurement Y_0 and a follow-up measurement Y_1 of log_{10} viral load, the outcome of interest Y was defined as follows: If Y_1 was above the lower limit of detection ($Y_1 > 1.7$), then $Y = Y_1 - Y_0$; if Y_1 was below the detectability limit, however, we imputed Y as the maximum decrease in viral load detected in the population, which was -4.2 log. Under this definition, both large drops in viral load from a high baseline and any achievement of an undetectable viral load (regardless of baseline) were treated as clinical successes. When several viral loads were measured between 4 and 36 weeks after regimen change, the first was used; duration from initiation of the salvage regimen until outcome measurement was included in the adjustment set W.

In summary, each TCE contained a baseline viral genotype, summarized in a vector **A** of binary variables defining the presence or absence of each of a list of candidate PI resistance mutations, a new antiretroviral regimen containing lopinavir initiated following the genotype, and an outcome Y capturing the change in log_{10} viral load at 4-36 weeks (measured before any subsequent changes in regimen) over baseline log_{10} viral load. In addition, each TCE contained a set W of adjustment variables, which included summaries of the non-PI mutations in the viral genotype, as well as covariates collected both prior to and following the genotype. We aimed to rank the candidate PI-mutations based on their impact on clinical outcome. In the sections that follow, we discuss several general approaches to research questions of this type, and discuss their implementation in the context of this data example.

3 Background: Statistical methods for biomarker discovery

3.1 Marginal vs. adjusted biomarker-outcome associations.

One straightforward approach to biomarker discovery is to assess the unadjusted association between each candidate biomarker and the outcome, or in other words, to estimate $E(Y|A = 1) - E(Y|A = 0)$ for each candidate A. In some settings the unadjusted association may be the quantity of interest, particularly when biomarkers can be experimentally manipulated. For example, if the researcher is able to induce specific mutations in a virus without altering other key covariates and then to compare viral replication in the presence and absence of each mutation, then assessment of marginal associations may be an appropriate approach.

In others settings, however, the marginal association between a candidate biomarker and the outcome can be misleading, or fail to capture the underlying mechanistic relationship of interest. When dealing with observational or clinical data, covariates are often present that are both associated with the candidate biomarker and also affect the outcome via a pathway independent of the biomarker. Such covariates are known in the epidemiologic literature as confounders.

The HIV data example illustrates how confounding of a biomarker effect can occur. HIV-infected patients with a given mutation may disproportionately include subjects

with an extensive treatment history. Because past treatment can strongly affect the presence of other mutations, past treatment patterns can cause a viral mutation with no effect on resistance to occur commonly with mutations that do strongly affect resistance. The candidate mutation may thus appear to confer resistance when in fact it is simply acting as a marker for past treatment history and the presence of other mutations. The picture is further complicated by the fact that in HIV infection, past mutations can be "archived" and remain present only in latent virus. Such archived mutations are not observable, but can still impact clinical response. We aimed to capture information about these archived mutations via covariates describing a subject's treatment history prior to initiation of the salvage regimen. In the HIV application, then, controlling for the presence of other mutations and for past treatment history allows us to isolate to what extent any decreased virologic response we observe is due to the presence of the candidate mutation being considered.

In the absence of residual confounding, the W-adjusted VIM $E_W(E(Y|A=1,W) - E(Y|A=0,W))$ corresponds to the mean causal effect of the biomarker on the outcome [13]. In the HIV example, if one is willing to assume that the measured covariates W are sufficient to control for confounding, adjustment can be used to estimate the causal effect of each candidate mutation on virologic response, defined as the mean difference in outcome that would have been observed if the researcher had somehow induced each mutation to be present versus absent in the entire study population. Depending on one's philosophy regarding causal effects, however, one may not be comfortable estimating the effect of a covariate on which one cannot intervene. Such a non-experimental scenario arises frequently in the context of biomarker discovery; it is often not possible, even theoretically, to "set" the level of a candidate biomarker and then to observe the change in outcome.

It also may not be possible to assume that all confounding is controlled for. Additional confounders may be unknown or simply unmeasured. In addition, even if the measured covariates W control adequately for confounding, it will not be possible to adjust for all covariates W if there is insufficient variation, or experimentation, in the occurrence of the candidate biomarker within strata of W. For example, if a mutation *always* occurs among subjects with a specific treatment history, then there is not sufficient information in the data to estimate the difference in clinical response that would be seen in the presence versus absence of the mutation in this sub-population. In the data example, the candidate PI mutations were highly collinear; as a result, for a given candidate mutation, we were unable to adjust for the presence of the other candidate PI mutations.

When estimation of the causal effect of a candidate biomarker is not feasible, adjustment of the association between biomarker and outcome for a set of covariates W often remains desirable. The quantity $E(Y|A=1,W=w) - E(Y|A=0,W=w)$ is interpretable as the difference in mean outcome in the presence versus absence of the candidate biomarker among subjects or observations with the same values of all covariates ($W=w$), and the VIM is simply the mean of these differences with respect to the empirical distribution of W. Adjustment for covariates W may be desirable as a means to reduce (rather than eliminate) the dependence of the biomarker-outcome association on the confounding structure of the data, resulting in a parameter that comes closer to

reflecting an underlying mechanistic relationship of interest. In addition, unlike unadjusted associations, the W-adjusted VIM $E_W(E(Y|A=1,W)-E(Y|A=0,W))$ does not depend on the joint distribution of A and W, and can thus provide more robust findings when applied to populations with similar marginal distributions of W but distinct confounding structures. For example, populations where antiretroviral treatment has been used differently in the past may have different relationships between a candidate protease resistance mutation and the mutations present in other viral enzymes. Controlling for past treatment and the presence of other mutations aims to improve the chances that protease mutations identified as important to virologic response in the current dataset will remain important in future treatment settings.

3.2 Adjustment for post-biomarker covariates.

Selecting which covariates to adjust for when estimating the VIM requires careful thought and substantial background knowledge about the specific data application to which the method is being applied. We discussed above the need in the HIV data example to control for at least two types of baseline covariates, treatment history prior to salvage regimen initiation and the presence of non-PI mutations. However, in some settings it may also be desirable to adjust for covariates that occur after, and may be affected by, the candidate biomarker of interest.

In the HIV data example, the non-PI drugs contained in the salvage regimen, assigned after assessment of viral genotype, may differ according to the presence of a candidate mutation. Such a scenario could arise, for example, if the clinician observed a mutation known to result in high-level resistance, and in response increased the potency of the subject's background (non-PI) regimen. To the extent that differences in background regimen impact clinical response, they have the potential to obscure drug resistance caused by the candidate mutation. In the causal inference framework, this scenario can be viewed as a (spurious) indirect effect of the mutation. Our aim is to estimate the direct effect of the mutation on clinical response, blocking any possible effect the presence of the mutation might have on the clinician's choice of background salvage regimen.

One option is to simply include post-biomarker covariates together with baseline covariates in the covariate set W. However, interpretation of the resulting W-adjusted VIM requires careful thought in the context of the specific data example to which it is being applied. Let W_b denote baseline covariates (occurring prior to the biomarker A), and let Z denote covariates occurring after, and affected by, A. At an individual level, the quantity $E(Y|A=1,Z=z,W_b)-E(Y|A=0,Z=z,W_b)$ corresponds (under assumptions on confounders - see [10]) to the effect of the biomarker on the outcome holding the intermediate variables Z at a fixed level. The mean of these individual effects provides a population summary: $E_W(E(Y|A=1,Z=z,W_b)-E(Y|A=0,Z=z,W_b))$. In the HIV example, this quantity would correspond with estimating the mean difference in virologic response if the researcher induced a candidate mutation to be present versus absent, and assigned a salvage regimen with fixed characteristics regardless of the presence of the mutation.

If one is willing to assume the absence of interaction between A and Z, then

$$E_{W_b}(E(Y|A=1,Z=z,W_b) - E(Y|A=0,Z=z,W_b)) \\ = E_{ZW_b}(E(Y|A=1,W_b,Z) - E(Y|A=0,W_b,Z)). \quad (1)$$

In other words, averaging over the empirical distribution of the post-biomarker covariates, Z, will not alter the estimated VIM, and thus the direct effect of interest can be estimated by simply including post-biomarker covariates together with baseline covariates in the adjustment set W. In the HIV example, the no-interaction assumption corresponds with assuming that the effect (or adjusted VIM) for each candidate PI mutation does not differ depending on the characteristics of the background regimen, a reasonable assumption given that PI mutations are not expected to affect response to non-PI drugs. In the analyses reported, characteristics of the (non-PI) background regimen were therefore included in the adjustment set W.

An additional common post-biomarker covariate is the duration between assessment of the biomarker and measurement of the outcome. To the extent that this duration is variable, differs depending on the presence of the biomarker, and affects the outcome, it has the potential to obscure the VIM of interest. In the HIV example, the outcome viral load was assessed between 4 and 36 weeks following salvage regimen initiation, and viral loads observed sooner following salvage initiation were likely to be higher. If the presence of a candidate mutation affected the time at which viral load was monitored, duration until the outcome was monitored could thus serve as an additional source of a spurious indirect effect. In the analyses reported in this article, time until viral load assessment was included as a covariate in W, according to the following rationale: 1) If the presence of the candidate mutation did not affect duration until outcome assessment, this duration could not serve as a source of an indirect effect, and inclusion of duration as a covariate did not require any additional assumptions; however, given the association between duration and the outcome, the inclusion of this covariate would be expected to improve efficiency. 2) If the presence of the candidate mutation did affect duration until outcome assessment, we wished to control for this indirect effect; inclusion of duration as a covariate allowed us to do this, again under the no interaction assumption (interpretable in this case as assuming that the effect of the mutation on virologic response did not vary over time). We note that inclusion of duration until outcome assessment is one possible way to address a potentially informative censoring mechanism; alternatives, such as the use of inverse probability weights [14], are beyond the scope of this article.

In summary, depending on the data application, inclusion of post-biomarker covariates in the adjustment set W may be warranted. However, such a decision requires careful consideration of the interpretation of the resulting W-adjusted VIM. In the following section, we return to the estimation of this parameter.

3.3 A traditional approach to the estimation of variable importance measures.

A common approach to the estimation of W-adjusted VIMs focuses on estimation of the conditional expectation $E(Y|A,W)$ of the outcome given the biomarker and covari-

ates, using standard maximum likelihood estimation. Given an estimate of $E(Y|A,W)$, the VIM can be estimated by simply evaluating this object at the values $A = 0$ and $A = 1$, and averaging the resulting differences across the population. Such an approach of intervening on the likelihood corresponds to the G-computation formula of Robins [9], applied in the setting of a single time-point. Frequently, the number of covariates W is large and the functional form of $E(Y|A,W)$ is unknown. Multiple algorithms are available to learn this form data-adaptively; examples include classification and regression trees [3], random forests [2], least angle regression [5], and the Deletion/Substitution/Addition (D/S/A) algorithm [12]. Either cross-validation or some form of penalization of the likelihood are generally used to select the level of model complexity providing the optimal bias-variance trade-off for the purposes of prediction; in the case that Y is continuous, this corresponds to selecting the level of complexity which minimizes the mean squared error.

Such an approach is appropriate if the goal of the analysis is to find the optimal predictor of the outcome Y given A and W. However, biomarker discovery often aims instead to evaluate a list of candidate biomarkers, rank them in terms of importance, and identify those significantly associated with the outcome. When the goal of analysis is to estimate the W-adjusted VIM for each of the candidate biomarkers, a different estimation approach may be warranted. To understand why, consider the HIV data example.

The number of covariates in this application, as in many biomarker applications, is very large, consisting of multiple mutations, salvage regimen characteristics, baseline characteristics of the subject such as viral load and CD4 count, and the subject's past antiretroviral treatment experience. A conventional approach would attempt to choose the model that best predicts virologic response as a function of the candidate mutation and these covariates. Given the large number of covariates, a reasonable approach would be to apply some data-adaptive regression algorithm to select this model. However, standard data-adaptive approaches aim to achieve the optimal bias-variance tradeoff for the entire conditional expectation of Y given A and W. Because the VIM is a much smoother parameter, a model fit for the purpose of prediction will generally not provide the best bias-variance trade-off for the purpose of estimating the VIM. Furthermore, a predictor constructed using conventional methods is likely to involve multiple terms that do not contain the candidate mutation; for example, baseline viral load and CD4 T cell count are likely to make important contributions to virologic response regardless of mutation profile. Mis-specification of such terms in, for example, a traditional multivariable regression model can result in bias in the estimated effect of the mutation, even under the null hypothesis of no mutation effect.

In summary, in the context of biomarker discovery, prediction is often not the underlying goal of analysis. Traditional approaches invest in achieving a good fit for the entire conditional expectation of Y given A and W; however such a fit is not targeted at the biomarker-specific VIM of interest. In contrast, *targeted* maximum likelihood estimation of the VIM, introduced in the following section, allows the researcher to focus on the importance of each mutation in turn, reducing bias in the adjusted VIM estimate and improving robustness to mis-specification of the model for $E(Y|A,W)$.

4 Targeted maximum likelihood estimation.

In this section, we provide a practical overview of targeted maximum likelihood estimation of variable importance measures. The formal statistical theory behind targeted maximum likelihood has been published elsewhere [15]. Here, our aim is to make this material practically accessible to the practitioner who wishes to apply targeted maximum likelihood estimation to improve biomarker discovery.

The density of the observed data $O = (W, A, Y)$ is defined by the marginal distribution of covariates W, the conditional distribution $P(A|W)$ of the biomarker given covariates, and the conditional distribution $P(Y|A, W)$ of the outcome Y given A and W. Unlike standard approaches to VIM estimation (which rely entirely on estimating $E(Y|A, W)$), targeted maximum likelihood estimation also involves estimation of $P(A|W)$. This estimate of the conditional distribution of the biomarker given covariates is used to update an initial estimate of $E(Y|A, W)$ in such a way that evaluating the updated estimate at $A = 1$ and $A = 0$ and taking the empirical mean results in an estimator of the W-adjusted VIM with reduced bias and improved robustness to model mis-specification.

Denote our parameter of interest, the W-adjusted VIM, by

$$\theta \equiv E_W \Big[E(Y|A = 1, W) - E(Y|A = 0, W) \Big]. \tag{2}$$

To ensure that this parameter is well-defined, we will assume that

$$0 < P(A = 1|W) < 1 \tag{3}$$

with probability one, or in other words, that some variation in the biomarker exists within each stratum of W.

We first summarize the basic steps involved in targeted maximum likelihood estimation of θ before going on to discuss each in detail, illustrated in the context of the data example. Implementation of the targeted maximum likelihood involves the following steps:

1. Estimating the conditional expectation of Y given A and W. We denote this initial estimate $Q_n^0(A, W)$.
2. Estimating the conditional distribution of the biomarker given covariates. We denote this estimate $g_n^0(A, W)$.
3. For each subject, calculating a specific covariate, based on the subject's observed values for A and W and using the estimate $g_n^0(A, W)$. We denote this covariate $h(A, W)$.
4. Updating the initial regression $Q_n^0(A, W)$ by adding the covariate $h(A, W)$ and estimating the corresponding coefficient by maximum likelihood, holding the remaining coefficient estimates fixed at their initial values. We denote this updated regression $Q_n^1(A, W)$.
5. Evaluating the updated regression at $A = 1$ and $A = 0$ to get two predicted outcomes for each subject and taking the empirical mean of the difference across the population to obtain a targeted estimate of the VIM.

4.1 An initial estimate of $E(Y|A,W)$.

The first step in targeted maximum likelihood estimation consists of obtaining an initial estimate of the conditional expectation $E(Y|A,W)$ of Y given A and W, as one would do in a standard G-computation approach to variable importance estimation. The number of covariates W will often be large, and the functional form for $E(Y|A,W)$ will often be unknown. In this case, as discussed in Section (3.3), a range of data-adaptive approaches are available to obtain an estimate $Q_n^0(A,W)$.

In the HIV data example, we were faced with a large number of candidate covariates, detailed in Section 2.2. These included mutations other than the candidate mutation of interest (incorporated both as individual covariates and summarized using measures such as drug-specific susceptibility scores), various summaries of past treatment history, baseline laboratory data on CD4 T cell count and viral load, time until outcome assessment, and summary measures of the background regimen and its estimated activity given baseline genotype. To reduce the size of the adjustment set W, we first performed a dimension reduction based on the unadjusted association of each candidate covariate with the outcome Y; the covariates with the 50 smallest p-values were retained.

Following this dimension reduction, we applied the D/S/A algorithm [12] to obtain an initial estimate $Q_n^0(A,W)$ based on the remaining 50 covariates. The D/S/A algorithm is a data-adaptive algorithm for polynomial regression that generates candidate predictors as linear combinations of polynomial tensor products in continuous and/or binary covariates. These candidate estimators are indexed by the number and complexity of the terms, and the optimal candidate is selected using cross-validation. In estimating $E(Y|A,W)$, the D/S/A algorithm considered candidate estimators with up to two-way interaction terms and a maximum quadratic order for each term. Specifically, $E(Y|A,W)$ was modelled by first selecting a model for $E(Y|W)$ with a maximum of 10 terms, then adding the term A to the selected model, and finally re-running the algorithm to select a model for $E(Y|A,W)$, forcing previous terms to be in the model and allowing the D/S/A algorithm to add up to 5 new terms.

This initial estimate of $E(Y|A,W)$ was evaluated at $A=1$ and $A=0$, and the empirical mean of the difference was used to estimate VIMs according to the G-computation approach. In other words, the G-computation estimate of the VIM was given by

$$\theta_n^{G-comp} = \frac{1}{n}\sum_{i=1}^{n} Q_n^0(1,W_i) - Q_n^0(0,W_i). \tag{4}$$

The targeted maximum likelihood estimate of the VIM also made use of this initial estimate Q_n^0, updated according to the following steps.

4.2 Estimation of $P(A|W)$.

The next step in the targeted estimation of VIMs consists of estimating the conditional distribution of A given W. In the current application, A is binary so that a logistic regression model can be used for this purpose. In fitting such a model, we first employed the same dimension reduction on W as used in fitting $E(Y|A,W)$. We then

used the D/S/A algorithm to data-adaptively select an appropriate logistic regression model for the probability of having the candidate mutation given W. The D/S/A algorithm was run with a maximum of two-way interactions, a maximum quadratic order for each term, and a maximum of ten terms. The practical performance of the targeted maximum likelihood estimator can be improved somewhat by ensuring that no estimated treatment probabilities $g_n^0(A,W)$ are very close to zero; here, we do so by setting estimated treatment probabilities smaller than 0.01 to 0.01.

4.3 Calculation of $h(A,W)$ and update of $Q_n^0(A,W)$.

Using the resulting estimate $g_n^0(A,W)$, the next step is to calculate the following covariate, denoted $h(A,W)$, for each subject:

$$h(A,W) \equiv \left(\frac{I(A=1)}{g_n^0(1,W)} - \frac{I(A=0)}{g_n^0(0,W)} \right). \tag{5}$$

A one-step adjustment to the initial regression estimate $Q_n^0(A,W)$ is performed by adding the covariate $h(A,W)$ to this regression and obtaining a maximum likelihood estimate ϵ_n of the corresponding coefficient ϵ, holding all other coefficient estimates fixed at their initial values. The estimate ϵ_n can thus be obtained by regressing Y on $h(A,W)$ using $Q_n^0(A,W)$ as an offset. The updated estimate $Q_n^1(A,W)$ is then given by

$$Q_n^1(A,W) = Q_n^0(A,W) + \epsilon_n h(A,W). \tag{6}$$

The corresponding targeted estimate of the marginal VIM is given by

$$\theta_n^{T-MLE} = \frac{1}{n} \sum_{i=1}^{n} Q_n^1(1,W_i) - Q_n^1(0,W_i). \tag{7}$$

The targeted maximum likelihood estimator is thus identical to the G-computation estimator described above except that it is based on the updated regression fit $Q_n^1(A,W)$ rather than the initial fit $Q_n^0(A,W)$.

4.4 Advantages of targeted maximum likelihood estimation.

Standard approaches to the estimation of variable importance rely entirely on the estimation of the conditional expectation of the outcome given the biomarker and covariates. The approach presented here provides a means to target this regression estimate specifically at the parameter of interest (in this case the W-adjusted VIM). In the context of the HIV data, for example, targeted maximum likelihood estimation of W-adjusted variable importance allows us to obtain a targeted estimate of the significance of each candidate resistance mutation in turn.

If the initial estimate of $E(Y|A,W)$ is based on standard multivariable or logistic regression, implementing the targeted maximum likelihood estimator is simply a matter of adding a covariate to the initial regression and estimating the corresponding

coefficient by maximum likelihood. The result of this single-step adjustment is a reduction in bias for the parameter of interest [15]. In addition, the targeted VIM estimate has improved robustness to model mis-specification in comparison to a G-computation estimate based on the initial regression fit. Specifically, the G-computation estimator is consistent only if the model for $E(Y|A,W)$ is correctly specified. In contrast, the targeted maximum likelihood estimator is consistent if the model for *either* $E(Y|A,W)$ *or* $P(A|W)$ is correctly specified. This added robustness is particularly valuable in contexts where the dependence of the biomarker on covariates is easier to model than the dependence of the outcome on biomarker and covariates.

Standard errors estimates and p-values for the targeted maximum likelihood VIM estimator can be obtained using the non-parametric bootstrap. This approach provides a straightforward means to address dependence between observations, as occurred in the data example because a single subject could contribute more than one TCE to the analyses. The non-parametric bootstrap also offers an opportunity to perform resampling-based approaches to multiple testing without substantial additional computer time.

5 Results: Identification of HIV mutations associated with decreased viral susceptibility to lopinavir.

In this section, we present the results of applying three different approaches to assess the importance of each of a set of candidate PI mutations in determining clinical virologic response to lopinavir:

1. Estimation of the unadjusted association $E(Y|A=1) - E(Y|A=0)$, based on univariate regression of Y on A.
2. Estimation of the W-adjusted VIM $E_W(E(Y|A=1,W) - E(Y|A=0,W))$, based on the G-computation estimator (4).
3. Estimation of the W-adjusted VIM $E_W(E(Y|A=1,W) - E(Y|A=0,W))$, based on the targeted maximum likelihood estimator (7).

Four hundred and one TCEs among 372 subjects involved initiation of a salvage regimen containing lopinavir and met all of our inclusion criteria. The frequency of the various candidate PI mutations among these TCEs is summarized in Table 1. Here and subsequently, mutations are denoted by the position of the change in the HIV protease enzyme, followed by a letter indicating the amino acid that has been substituted (e.g. 53LY refers to a substitution of leucine or tyrosine at protease position 53). As discussed in Section 3 and stated formally in equation (3) in Section 4, adjustment for covariates W requires that there be variation in the presence of the biomarker within strata of W. In order to help ensure sufficient variation and the ability to control adequately for confounding, we estimated VIMs only for those mutations which occurred in at least 20 TCEs; among the mutations that had to be excluded based on this criterion are the important lopinavir resistance mutations 50V, 84C, and 88S. In addition, we assessed the extent of variation among the remaining mutations by examining the fitted probabilities $g_n^0(A,W)$. For a few of these mutations, most notably 54LMST

Table 1: Frequency of candidate protease inhibitor mutations among the 401 TCEs included in the analysis. *VIMs were estimated only for those mutations which occurred in at least 20 TCEs. For those mutations present in at least 20 TCEs, % Violations gives the percentage of TCEs with fitted mutation probabilities* < 0.05 *or* > 0.95*; a high percentage may reflect a lack of variation in the distribution of the mutation that can lead to unreliable VIM estimates.*

Mutation	Frequency	% Violations
10FIRVY	217	3%
16E	9	–
20IMRTVL	115	0%
23I	4	–
24IF	16	–
30N	45	64%
32A	0	–
32I	21	58%
33F	44	51%
36ILVTA	141	0%
46ILV	143	0%
47V	17	–
48VM	16	–
48AST	1	–
50V	5	–
50L	0	–
53LY	33	0%
54LMST	36	84%
54VA	84	0%
63P	311	5%
71TVI	181	0%
73CSTA	66	35%
82AFST	100	6%
82MLC	4	–
84AV	73	28%
84C	2	–
88DTG	44	36%
88S	9	–
90M	171	0%

and 30N, a high proportion of the fitted probabilities were less than 0.05 or greater than 0.95, suggesting that they may not exhibit enough variation within strata of W to allow for reliable VIM estimation. The results presented for these mutations should thus be interpreted with care.

It was not clear based on background knowledge whether the presence of mutations affected the duration until the outcome viral load was measured. We investigated this potential dependence by using box plots to compare the distribution of outcome monitoring times in the presence versus absence of each mutation. These plots did not suggest any major differences in the distribution of monitoring times according to the presence or absence of any mutation. In addition, we fit a data-adaptive model of the conditional hazard of viral load monitoring over time in order to examine the potential dependence of monitoring on the presence of candidate mutations and baseline covariates. The data-adaptively selected model included as single covariate the time that had elapsed since initiation of the new treatment regimen. Together, these findings suggest that the presence of particular mutations did not strongly affect monitoring time, reducing concern regarding the assumption that mutation effect was constant over time (discussed in Section 3.2).

Table 2: Estimated VIMs and associated p-values for candidate PI mutations. *Score refers to the resistance score assigned to a mutation by the Stanford HIVdb scoring system (accessed on 7/18/2006).*

Mutation	**Score**	**Unadjusted**		***G*-comp**		**T-MLE**	
		VIM	*p-value*	*VIM*	*p-value*	*VIM*	*p-value*
10FIRVY	2	0.56	< 0.01	0.28	0.12	0.26	0.30
20IMRTVL	2	0.46	0.02	0.39	0.04	0.37	0.04
30N	0	-1.09	< 0.01	-0.60	0.03	-0.20	0.72
32I	10	0.80	0.01	0.63	0.03	0.81	< 0.01
33F	5	0.83	< 0.01	0.49	0.05	1.12	0.02
36ILVTA	1	0.29	0.10	0.39	0.03	0.39	0.04
46ILV	11	0.44	0.01	0.18	0.32	0.13	0.60
53LY	3	0.54	0.04	0.32	0.28	0.32	0.33
54LMST	10	0.67	0.01	0.15	0.55	0.16	0.72
54VA	11	0.86	< 0.01	0.69	< 0.01	0.61	< 0.01
63P	2	0.10	0.57	-0.02	0.90	-0.07	0.72
71TVI	2	0.34	0.03	0.24	0.13	0.24	0.17
73CSTA	2	0.79	< 0.01	0.61	0.02	0.46	0.36
82AFST	20	0.68	< 0.01	0.49	0.02	0.64	< 0.01
84AV	11	0.50	0.02	0.25	0.19	0.49	0.04
88DTG	0	-0.86	< 0.01	-0.50	0.05	-0.37	0.33
90M	10	0.52	< 0.01	0.45	0.02	0.45	0.02

Table 2 summarizes the unadjusted associations and estimates of the W-adjusted VIM based on the G-computation and targeted approaches, along with associated p-

Table 3: Candidate PI mutations ranked according to the p-values of three distinct VIM estimates. *Score refers to the resistance score assigned to a mutation by the Stanford HIVdb scoring system (accessed on 7/18/2006). Mutations marked with an asterisk have a negative VIM estimate, suggesting that they contribute to an improved rather than diminished virologic response.*

Score		**Unadjusted**		**G-comp**		**T-MLE**	
Mutation	*Score*	*Mutation*	*p-value*	*Mutation*	*p-value*	*Mutation*	*p-value*
82AFST	20	30N*	< 0.001	54VA	< 0.001	82AFST	0.001
54VA	11	54VA	< 0.001	82AFST	0.018	54VA	0.003
46ILV	11	82AFST	< 0.001	90M	0.019	32I	0.003
84AV	11	33F	< 0.001	73CSTA	0.019	90M	0.024
90M	10	10FIRVY	0.001	32I	0.033	33F	0.024
32I	10	73CSTA	0.001	30N*	0.033	36ILVTA	0.035
54LMST	10	88DTG*	0.001	36ILVTA	0.034	84AV	0.037
33F	5	90M	0.003	20IMRTVL	0.043	20IMRTVL	0.039
53LY	3	32I	0.014	33F	0.051	71TVI	0.174
10FIRVY	2	46ILV	0.015	88DTG*	0.051	10FIRVY	0.301
73CSTA	2	54LMST	0.015	10FIRVY	0.123	53LY	0.330
20IMRTVL	2	84AV	0.016	71TVI	0.130	88DTG*	0.330
71TVI	2	20IMRTVL	0.016	84AV	0.193	73CSTA	0.361
63P	2	71TVI	0.034	53LY	0.277	46ILV	0.600
36ILVTA	1	53LY	0.039	46ILV	0.321	63P*	0.719
30N	0	36ILVTA	0.097	54LMST	0.551	30N*	0.719
88DTG	0	63P	0.574	63P*	0.898	54LMST	0.719

value. Table 3 shows three different rankings for the set of candidate mutations, based on the p-values generated by each of the three approaches. The mutation ranking generated by the current Stanford scoring system is included for comparison. Inference was based on non-parametric bootstrap sampling, respecting the subject rather than the TCE as the independent unit of analysis. The resulting p-values were adjusted for multiple testing using the Benjamini-Hochberg method [1] to control the false discovery rate (aiming to ensure that the expected proportion of false positives was 0.05).

Among the 17 candidate PI mutations considered here, the Stanford scoring system identifies the following seven mutations as major contributors to lopinavir resistance: 82AFST, 54VA, 46ILV, 84AV, 90M, 32I, and 54LMST; the remaining ten mutations are thought to make minor or no contributions to resistance. The unadjusted association analysis yielded significant p-values for all but two of the candidate PI resistance mutations (36ILV and 63P). The significant subset thus included eight mutations thought to have a minor or no effect on lopinavir resistance. Among these were the mutations 30N and 88DTG, both estimated to be significantly protective. The protective association of 30N with the outcome was in fact ranked the most important of the unadjusted

associations. In addition, multiple mutations considered by current knowledge to have only minor effects on resistance (for example, 33F, 10FIRV and 73CST) ranked higher than most of the known major lopinavir resistance mutations (such as 90M, 32I, and 54LMST).

After adjusting for covariates using G-computation, fewer mutations were identified as significant, and the resulting ranking agreed to a greater extent with current knowledge. Specifically, this approach identified eight mutations as having a significant impact on lopinavir resistance, with an additional two mutations found to be borderline significant (p-values of 0.051 for 33F and 88DTG). This group of ten mutations includes both four of the seven major lopinavir resistance mutations and six mutations thought to make minor or no contributions to resistance. In particular, we note that the mutations 30N and 88DGT were still identified as having a protective effect.

Targeted maximum likelihood estimation of the adjusted VIM provided the ranking in best agreement with current knowledge. The significant subset of mutations identified by this approach included five of the seven major known mutations, and only three minor mutations (33F, 36ILV, 20IMRTV). The mutation considered most important for lopinavir resistance, 82AFST, was ranked highest, followed by three major known lopinavir resistance mutations (32I, 54AV and 90M). Unlike G-computation, targeted maximum likelihood also identifies the major lopinavir resistance mutation 84AV as a significant contributor to resistance. In addition, unlike the other two approaches, it did not rank either 88DGT or 30N as significantly protective. Two mutations thought to be important for lopinavir resistance, 46ILV and 54LMST, were not identified by targeted VIM estimation. However, Table 1 shows that for the mutation 54LMST, 84% of observations had fitted mutation probabilities < 0.05 or > 0.95, suggesting a lack of variation in 54LMST within strata of W that may lead to unreliable VIM estimates. In addition, *in vitro* experiments examining the effect of 46ILV on viral phenotype suggest that this mutation may in fact be less important for lopinavir resistance than previously thought [8].

6 Discussion.

6.1 HIV resistance mutations.

The current article discussed how targeted maximum likelihood estimation of variable importance measures can be used in biomarker discovery. Motivation for the method, details of its implementation, and interpretation of results were illustrated using an example from the treatment of HIV infection. We estimated the importance of each of a set of candidate PI mutations for clinical virologic response to treatment with the commonly used PI drug lopinavir, adjusted for covariates including treatment history, the presence of non-PI mutations, and characteristics of the background regimen.

Our analysis suggests that targeted maximum likelihood estimation of VIM represents a promising new approach for studying the effects of HIV mutations on clinical virologic response to antiretroviral therapy. The subset of mutations identified by this approach as significant contributors to lopinavir resistance was in better agreement with current knowledge than the subsets identified by an unadjusted analyses or the

G-computation approach. Specifically, the unadjusted analysis identified as significant all but two of the candidate mutations, including eight mutations thought to have a minor or no effect on lopinavir resistance. G-computation reduced the significant subset to four of the seven mutations thought to make major contributions to lopinavir resistance, while still including six mutations thought to make only a minor or no contribution to resistance. In contrast, the significant subset of mutations identified by targeted maximum likelihood included five of the seven major known mutations and only three minor mutations. In addition, the specific ranking provided by targeted VIM estimation also agreed better with current understanding than did the rankings generated with alternative methods.

While targeted VIM estimates were able to replicate most known findings, they also suggested that the mutation 46ILV may be less important in determining resistance to lopinavir than previously thought. As mentioned in Section 5, this finding has some support from *in vitro* studies [8], suggesting that a more detailed investigation of the role of this mutation may be warranted. Taken as a whole, the promising results reported here suggest that further application of the targeted VIM approach may result in improvements to existing genotypic interpretation algorithms.

6.2 Targeted maximum likelihood.

As illustrated in this article, targeted maximum likelihood estimation offers an improvement in robustness over conventional likelihood-based approaches that is straightforward to implement using standard statistical software. Specifically, the approach remains consistent if we mis-specify how virologic response depends on the mutation and all covariates, but correctly model how the presence of the mutation depends on covariates. The resulting targeted VIM estimates provide a means to both rank candidate biomarkers and to identify a subset of biomarkers as relevant for a given outcome. The current article focused primarily on VIM for a continuous outcome. Generalization to a binary outcome modelled using logistic regression is straightforward, as was mentioned briefly. The method can further be generalized to alternative approaches for obtaining an initial estimate of $E(Y \mid A, W)$.

The double robust variable importance estimator introduced by van der Laan [13] provides similar advantages to the targeted VIM estimate in terms of improved robustness to model mis-specification. However, the targeted approach has several practical advantages. Many practitioners are more familiar with regression-based approaches, as used by the targeted estimator, than with the estimating function methodology employed by the double robust estimator. In addition, the targeted maximum likelihood VIM estimator can in many cases be implemented using standard software, in a natural extension of common regression approaches. These practical advantages, together with the improvement in robustness, make targeted maximum likelihood estimation of variable importance a promising new approach to biomarker discovery.

References

[1] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statstical Society: Series B*, 57:289–300, 1995.

[2] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[3] L. Breiman, J. H. Friedman, R. Olshen, and C. J. Stone. *Classification and regression trees*. The Wadsworth statistics/probability series. Wadsworth International Group, 1984.

[4] F. Brun-Vezinet, D. Descamps, A. Ruffault, B. Masquelier, V. Calvez, G. Peytavin, F. Telles, L. Morand-Joubert, J. L. Meynard, M. Vray, D. Costagliola, and Narval ANRS 088 Study group. Clinically relevant interpretation of genotype for resistance to abacavir. *AIDS*, 17(12):1795–1802, 2003.

[5] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least Angle Regression. *The Annals of Statistics*, 32(2):407–499, 2004.

[6] S. Hammer, M. Saag, M. Schechter, J.S. Montaner, R.T. Schooley, D.M. Jacobsen, M.A. Thompson MA, C.C. Carpenter, M.A Fischl, B.G. Gazzard, J.M. Gatell, M.S. Hirsch, D.A. Katzenstein, D.D. Richman, S. Vella, P.G. Yeni, P.A. Volberding, and International AIDS Society USA panel. Recommendations of the International AIDS Society - USA Panel. *Journal of the American Medical Association*, 296(7):827–843, 2006.

[7] K. Van Laethem, K. Van Vaerenbergh, J.C. Schmit, S. Sprecher, P. Hermans, V. De Vroey R. Schuurman T. Harrer, M. Witvrouw, E. Van Wijngaerden, L. Stuyver, M. Van Ranst, J. Desmyter, E. De Clercq, and A.M. Vandamme. Phenotypic assays and sequencing are less sensitive than point mutation assays for detection of resistance in mixed hiv-1 genotypic populations. *Journal of the Acquired Immunodeficiency Syndrome*, 22(2):107–118, 1999.

[8] S.-Y. Rhee, J. Taylor, G. Wadhera, A. Ben-Hur, D. L. Brutlag, and R. W. Shafer. Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences*, 103:17355–17360, 2006.

[9] J.M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy survivor effect. *Mathematical Modelling*, 7:1393–1512, 1986.

[10] J.M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3:143–155, 1992.

[11] R.W. Shafer, D.R. Jung, and B.J. Betts. Human immunodeficiency virus type 1 reverse transcriptase and protease mutation search engine for queries. *Nature Medicine*, 6(11):1290–1292, 2000.

[12] S.E. Sinisi and M.J. van der Laan. Deletion/Substitution/Addition algorithm in learning with applications in genomics. *Statistical Appliations in Genetics and Molecular Biology*, 3(1):Article 18, 2004. URL `www.bepress.com/sagmb/vol3/iss1/art18`.

[13] M.J van der Laan. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1):Article 2, 2006. URL `http://www.bepress.com/ijb/vol2/iss1/2`.

[14] M.J. van der Laan and J.M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer Series in Statistics. Springer Verlag, 2003.

[15] M.J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):Article 11, 2006. URL `http://www.bepress.com/ijb/vol2/iss1/11`.

7.3 *Data-adaptive Selection Of The Adjustment Set in Variable Importance Estimation*

The following article appears as it was originally published on the University of California, Berkeley Division of Biostatistics Working Paper Series website in 2008, http://www.bepress.com/ucbbiostat/paper231/.

Data-adaptive selection of the adjustment set in variable importance estimation

Oliver Bembom[1], W. Jeffrey Fessel[3], Robert W. Shafer[2], and Mark J. van der Laan[1]

[1]Division of Biostatistics, University of California, Berkeley, CA
[2]Division of Infectious Diseases,Center for AIDS Research, Stanford University, Palo Alto, CA
[3]Clinical Trials Unit, Kaiser Permanente, San Francisco, CA

Abstract

If estimates of the effect of a treatment variable on an outcome of interest are to be adjusted for a set of possible confounding factors, it is necessary to rely on the assumption of experimental treatment assignment (ETA) according to which each experimental unit has positive probability of being observed at any of the possible levels of the treatment variable regardless of the values the confounding factors may take on. Even if this assumption is only practically violated in the sense that certain values of the confounding factors cause some treatment levels to become not impossible, but at least highly unlikely, the adjusted variable importance parameter often becomes poorly identified in finite samples.

We introduce an algorithm that is intended to make variable importance estimation more robust with respect to violations of the ETA assumption. Two different identifiability criteria are proposed for deciding when an adjusted variable importance parameter cannot be reliably estimated from the data. These criteria are then used to identify a maximal set of adjustment variables for which the ETA assumption appears reasonably well satisfied. A more efficient estimator of the parameter corresponding to the selected adjustment set is then sought by selecting from among estimators making use of even smaller adjustment sets by minimizing an estimate of the mean squared error for the selected parameter.

A simulation study aimed at evaluating the benefits of this latter step suggests that it can lead to efficiency gains on the order of 100% if the ETA assumption is violated to some extent and to efficiency gains on the order of 35% if the ETA assumption is well approximated. The proposed algorithm is applied to the problem of identifying mutations in the protease enzyme of HIV that have an effect on virologic response to the commonly used antiretroviral drug lopinavir. While both unadjusted and fully adjusted analyses yield unsatisfactory results, the subset of significant mutations identified by the algorithm introduced here includes eight of the 12 known major lopinavir resistance mutations as well as two mutations that are thought to increase susceptibility to lopinavir. Two of the four major mutations not identified in our analysis occurred very rarely in our data set, giving the algorithm low power to detect any impact on virologic response. Recent in vitro experiments suggest that the other two major mutations not identified here may in fact be less important in determining lopinavir resistance than previously thought. The excellent agreement of the results reported here with current understanding of lopinavir resistance suggest that variable importance estimation based on data-adaptive selection of the adjustment set represents a promising new approach for studying the effects of HIV mutations on clinical virologic response to antiretroviral therapy as well as for biomarker discovery in general.

1 Introduction

Many applications in modern biology measure a large number of genomic or proteomic covariates and are interested in assessing the impact of each of these covariates on a particular outcome of interest. In a study of HIV-positive patients, for example, a researcher may genotype the virus infecting each patient to ascertain the presence or absence of a large number of mutations, in the hope of identifying mutations that affect how a patient's plasma HIV RNA level (viral load) responds to a new drug regimen. Along with an estimate of the impact of each mutation on viral load, the researcher would generally like to have a measure of the statistical significance of these estimates in order to identify those mutations that are most likely to be genuinely related to the outcome. Such information could then be used to inform the decision of which drugs should be included in the regimen of a patient with a particular pattern of mutations.

The simplest way of assessing the impact of a particular mutation on viral load would be to compare the virologic response among patients whose virus has the mutation to that among patients whose virus does not. If we find that patients in the first group respond much more poorly to a particular drug regimen, a clinician might be inclined not to give this regimen to a new patient entering his office who has this mutation. Patients in the first group are, however, also quite likely to differ from those in the second group in terms of the remaining mutations as well as other measured clinical covariates. The mutation of interest may, for example, be very common among patients who have previously failed several similar drug regimens, making them far more likely to also fail the current one, but very rare among other patients. If the clinician's new patient comes from a population that differs from our original study population in that the mutation is not associated with having previously failed similar drug regimens, we might be wrong to conclude that the regimen under consideration would be a poor choice in this situation. Since the impact of the mutation of interest on viral load is confounded by the remaining mutations as well as other clinical covariates, such unadjusted estimates thus do not generalize to a new population in which the mutation of interest and the confounding factors are related to each other in a different way.

We might thus be interested in estimating the impact of a given mutation on viral load that is not due to associations of this mutation with any of the other measured covariates. Specifically, we might ask: What difference in virologic response would we observe if we could somehow give every patient in our study population the mutation interest, holding the remaining covariates fixed at their current values, as opposed to the scenario in which we give none of the patients this mutation, holding again other covariates fixed? Any observed difference could then not be due to differences of the two populations with regard to the remaining covariates and would thus be more likely to generalize to a new population in which the mutation of interest and the other covariates may be related to each other differently.

While such adjusted variable importance estimates are thus often more interesting than the corresponding unadjusted estimates, they also rely on an additional assumption in order to be identifiable from the collected data. Specifically, the assumption of experimental treatment assignment (ETA) requires that the adjustment variables cannot take on a set of values such that the group of patients corresponding to those values shows no variation in the mutation of interest. This assumption would be violated if, for example, there existed a second mutation that always occurred in concordance with the mutation of interest. Since we would never observe patients that exhibited each of the two mutations in the absence of the other one, it would be impossible to disentangle the individual effects of these two mutations, precluding us thus from estimating their impact on viral load adjusting for the other mutation.

More commonly, the set of adjustment variables may contain covariates that are not perfectly predictive of the mutation of interest, but that still determine the presence or absence of that mutation in a nearly deterministic fashion. A second mutation may, for example, be so strongly correlated with the mutation of interest that 99% of patients with this second mutation also exhibit the mutation of interest. In such instances, a substantial amount of data would be required before the adjusted variable importance of the mutation of interest could be estimated in any reliable way. In smaller samples, it could easily occur by chance that we observe no patients that are discordant for these two mutations, again precluding us from obtaining an adjusted variable importance estimate. To distinguish this scenario from the one described in the previous paragraph, we refer to it as a practical rather than a theoretical violation of the ETA assumption.

Under either of these two violations of the ETA assumption, the desired adjusted variable importance is not identifiable from the data at hand, making any estimates of this parameter unreliable and hard to

interpret. A practical ETA violation, for example, often causes such estimates to become unstable and highly variable. An analysis that under such circumstances still aims to rank mutations based on adjusted variable importance estimates is thus bound to lead to unsatisfying results. Suppose, for example, that a mutation with no impact on viral load is strongly correlated with a second mutation that itself has a considerable impact. The practical ETA violation caused by this correlation would likely lead to highly variable and thus statistically non-significant adjusted variable importance estimates for both mutations. In this case, more useful results could be obtained by turning to variable importance estimates that do not attempt to adjust for the other mutation. This approach would likely yield significant estimates for both mutations, allowing the investigator to conclude that at least one of these two mutations has an impact on viral load. While we would have to acknowledge that we cannot disentangle the individual contributions of the two mutations, such a qualified identification of two mutations would generally seem preferable to the conclusion drawn from a fully adjusted analysis, according to which neither mutation would seem important in determining viral load.

These considerations suggest that it would be useful to have a criterion that could give the investigator a sense of the extent to which the variable importance parameter corresponding to a proposed adjustment set is identifiable from the data at hand. If this criterion suggested that the parameter corresponding to the full adjustment set was not well identified, it could then also be used to identify a smaller, more workable adjustment set. In this paper, we propose two criteria that can be used for this purpose, one based on the diagnostic for ETA bias developed by Wang et al. (2006), and one based on closed-form asymptotic bias estimates proposed by Bembom and van der Laan (2007a). In addition, we propose an approach for defining a sequence of nested candidate adjustment sets that, in combination with a given identifiability criterion, can be used to select an appropriate adjustment set data-adaptively.

Even if the variable importance parameter corresponding to a particular adjustment set is identified reasonably well by the data at hand, it may be advantageous to base estimation of this parameter on an adjustment set that in fact excludes additional covariates.The adjustment set defining the parameter of interest may, for example, contain a covariate that is a good predictor of the mutation under consideration, but only a weak predictor of viral load. Such a covariate will often be only a weak confounder of the relationship between the mutation and viral load, but can still lead to a mild practical violation of the ETA assumption that would cause the variable importance estimator to become more variable. Not adjusting for this covariate could thus, at the price of a slight increase in bias, offer a considerable reduction in variability, thus leading to an overall reduction in mean squared error. In this paper, we propose an approach that, given an adjustment set defining the parameter of interest, can be used to evaluate whether such additional exclusions from the adjustment set can be expected lead to more efficient estimates of that parameter. For the sake of clarity, we will refer to the adjustment set defining the parameter of interest, as selected based on a given identifiability criterion, as the targeted adjustment set; the possibly smaller adjustment set used in estimating this parameter, on the other hand, will be referred to as the effective adjustment set. The effective adjustment set is thus nested in the targeted adjustment set, which in turn is nested in the full adjustment set.

To summarize, this paper proposes an algorithm for variable importance estimation that first selects a targeted adjustment set defining the parameter of interest before then selecting an effective adjustment set that will be used in the estimation of this parameter. While the first step is aimed at making adjusted variable importance estimation robust to violations of the ETA assumption, the primary goal of the second step is to optimize efficiency. The remainder of the paper is organized as follows. In the next section, we review the formal definition of adjusted variable importance parameters as well as several estimators that have been proposed for these parameters. In section 3, we describe two different identifiability criteria that can be used for selecting the targeted adjustment set. The following section introduces our proposal for selecting the effective adjustment set. The possible efficiency gains that can be achieved by data-adaptively selecting the effective adjustment set are examined in a simulation study in section 5. Both steps of the proposed algorithm are then studied in an applied data analysis in section 6 that is aimed at ranking mutations in the protease enzyme of HIV based on their impact on virologic response to antiretroviral regimens containing the protease inhibitor lopinavir. We close with a brief discussion of possible extensions to the methodology described here.

2 Variable importance parameters and estimators

We consider the common point-treatment data structure consisting of n i.i.d. copies of $O = (W, A, Y)$, where $W = (W_1, \ldots, W_d)$ denotes the collection of measured confounders, A gives the treatment variable, and Y is the outcome of interest. For now we assume that A is binary. We would ideally like to estimate the marginal variable importance θ of A on Y controlling for W:

$$\theta \equiv E\Big[E[Y|A=1,W] - E[Y|A=0,W]\Big]. \tag{1}$$

This parameter is identified by the data under the ETA assumption according to which we have with probability 1.0 that, for $a \in \{0,1\}$,

$$P(A=a|W) > 0. \tag{2}$$

If there exists a w_1 such that $P(W=w_1) > 0$ and $P(A=a|W=w_1) = 0$ for $a=0$ or $a=1$, we say that the ETA assumption is theoretically violated. If (2) holds but there exists a w_2 such that $P(W=w_2) > 0$ and $P(A=a|W=w_2) \approx 0$ for $a=0$ or $a=1$, we say that the ETA assumption is practically violated.

We are here interested in identifying a maximal subset W^t of W such that we have with probability 1.0 that, for $a \in \{0,1\}$,

$$P(A=a|W^t) > \epsilon > 0, \tag{3}$$

assuring that the W^t-specific ETA assumption is neither theoretically nor practically violated. This in turn guarantees that the marginal variable importance of A on Y controlling for W^t can be identified from the data.

To identify the subset W^t, we first define a sequence of nested candidate adjustment sets. Since violations of the ETA assumption are caused by covariates that are highly predictive of A, we define these candidate adjustment sets based on a ranking of the confounders by their squared sample correlation with A. Specifically, each candidate adjustment set $W(\delta)$ will contain the δd covariates in W that have the lowest squared sample correlations with A, $0 \leq \delta \leq 1$. For this purpose, let $\rho_n^2(W_j, A)$ denote the squared sample correlation between W_j and A and let $q(\delta)$ denote the δ quantile of $\rho_n^2(W_1, A), \ldots, \rho_n^2(W_d, A)$. Then we can define $W(\delta) = (W_j : \rho_n^2(W_j, A) \leq q(\delta))$ as the collection of confounders with squared sample correlations no greater than the δ quantile $q(\delta)$ of squared sample correlations. The marginal variable importance parameter $\theta(\delta)$ corresponding to a candidate adjustment sets $W(\delta)$ is given by

$$\theta(\delta) \equiv E\Big[E[Y|A=1,W(\delta)] - E[Y|A=0,W(\delta)]\Big]. \tag{4}$$

Several estimators of such marginal variable importance parameters have been proposed (van der Laan, 2006). These estimators can typically be written as functions of the two nuisance parameters $g(\delta)(A,W) \equiv P(A|W(\delta))$ and $Q(\delta)(A,W) \equiv E[Y|A,W(\delta)]$. Assume that we have available preliminary estimates $g_n(\delta)$ and $Q_n(\delta)$ of these nuisance parameters; $g_n(\delta)$ may, for example, be obtained through a logistic regression of A on $W(\delta)$. Two popular variable importance estimators are then given by the G-computation estimator

$$\theta_n^{G-comp}(\delta) = \frac{1}{n}\sum_{i=1}^{n} Q_n(\delta)(1,W_i) - Q_n(\delta)(0,W_i) \tag{5}$$

and the Inverse-Probability-of-Treatment-Weighted (IPTW) estimator

$$\theta_n^{IPTW}(\delta) = \frac{1}{n}\sum_{i=1}^{n} \Big[I(A_i=1) - I(A_i=0)\Big]\frac{Y_i}{g_n(\delta)(A_i,W_i)}. \tag{6}$$

The G-computation estimator yields valid estimates of $\theta(\delta)$ if $Q(\delta)$ is estimated consistently; the IPTW estimator instead relies on consistent estimation of $g(\delta)$.

Recently a targeted maximum-likelihood estimator of $\theta(\delta)$ has been proposed that gives consistent estimates as long as either $g(\delta)$ or $Q(\delta)$ is estimated consistently (van der Laan and Rubin, 2006). This estimator is identical to the G-computation estimator (5) except that it is based on an updated regression fit $Q_n^1(\delta)(A,W)$ rather than the initial fit $Q_n(\delta)(A,W)$. The updated estimate $Q_n^1(\delta)$ is obtained in a straightforward manner by adding the covariate

$$h(\delta)(A,W) = \left(\frac{I(A=1)}{g_n(\delta)(1,W)} - \frac{I(A=0)}{g_n(\delta)(0,W)}\right) \tag{7}$$

to the original regression fit and obtaining a maximum likelihood estimate $\epsilon_n(\delta)$ of the corresponding coefficient $\epsilon(\delta)$, holding all other coefficient estimates fixed at their initial values. The estimate $\epsilon_n(\delta)$ can thus be obtained by regressing Y on $h(\delta)(A,W)$ using $Q_n(A,W)$ as an offset. The updated regression fit $Q_n^1(\delta)(A,W)$ is then given by

$$Q_n^1(\delta)(A,W) = Q_n(\delta)(A,W) + \epsilon_n(\delta)h(\delta)(A,W). \tag{8}$$

The corresponding targeted maximum-likelihood estimate of $\theta(\delta)$ can be obtained as

$$\theta_n^{T-MLE}(\delta) = \frac{1}{n}\sum_{i=1}^{n} Q_n^1(\delta)(1,W_i) - Q_n^1(\delta)(0,W_i). \tag{9}$$

This estimator solves the estimating equation

$$0 = \frac{1}{n}\sum_{i=1}^{n} D^{DR}(O_i|g_n(\delta), Q_n^1(\delta), \theta(\delta)) \tag{10}$$

corresponding to the double robust estimating function

$$\begin{aligned} D^{DR}(O|g(\delta), Q(\delta), \theta(\delta)) \quad &= \quad \Big[I(A=1) - I(A=0)\Big]\frac{Y - Q(\delta)(A,W)}{g(\delta)(A,W)} + \\ &\quad Q(\delta)(1,W) - Q(\delta)(0,W) - \theta(\delta). \end{aligned} \tag{11}$$

Under regularity conditions, the targeted maximum likelihood estimator is therefore asymptotically linear if at least one of the two nuisance parameters $g(\delta)$ and $Q(\delta)$ is estimated consistently (van der Laan and Robins, 2003). If both nuisance parameters are estimated consistently, the influence curve of the estimator is given by

$$IC^{T-MLE}(O|g_0(\delta), Q_0(\delta), \theta_0(\delta)) = c^{-1}D^{DR}(O|g_0(\delta), Q_0(\delta), \theta_0(\delta)), \tag{12}$$

where $g_0(\delta 0)$, $Q_0(\delta)$, and $\theta_0(\delta)$ denote the true values of the corresponding parameters and the standardizing constant c is given by

$$c = -\frac{\partial}{\partial\theta(\delta)} E D^{DR}(O|g_0(\delta), Q_0(\delta), \theta(\delta))\Big|_{\theta(\delta)=\theta_0(\delta)} = 1 \tag{13}$$

The asymptotic linearity of $\theta_n^{T-MLE}(\delta)$ under these conditions,

$$\sqrt{n}(\theta_n^{T-MLE}(\delta) - \theta_0(\delta)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} IC^{T-MLE}(O_i|g_0(\delta), Q_0(\delta), \theta_0(\delta)) + o_p(1), \tag{14}$$

implies in particular that

$$\sqrt{n}\Big(\theta_n^{T-MLE}(\delta) - \theta_0(\delta)\Big) \Rightarrow N\Big(0, \sigma^2(\delta) = Var(IC^{T-MLE}(O|g_0(\delta), Q_0(\delta), \theta_0(\delta))\Big). \tag{15}$$

An estimate $\sigma_n^2(\delta)$ of $\sigma^2(\delta)$ can be obtained as the sample variance of $IC^{T-MLE}(O|g_n(\delta), Q_n^1(\delta), \theta_n^{T-MLE}(\delta))$, allowing us to construct an asymptotic 95% confidence interval for $\theta(\delta)$ as

$$\theta_n^{T-MLE}(\delta) \pm 1.96\sqrt{\frac{\sigma_n^2(\delta)}{n}}. \tag{16}$$

If $g(\delta)$ is estimated consistently but $Q(\delta)$ is not, inference based on this approach is conservative (van der Laan and Robins, 2003).

Since the nuisance parameter $g(\delta)$ appears in the denominator of the covariate $h(\delta)(A,W)$, the targeted maximum-likelihood estimator can become quite unstable if some of the estimated treatment probabilities are close to zero, i.e. if the ETA assumption is practically violated. Its practical performance can often be improved somewhat by selecting a small value ϵ such as $\epsilon = 0.01$ and setting estimated treatment probabilities $g_n(\delta)(A,W) < \epsilon$ equal to ϵ.

We may also obtain a more stable estimator of $\theta(\delta)$ if we are willing to assume that the regression function $E[Y|A,W(\delta)]$ does not contain any interactions between A and $W(\delta)$. In that case, we have that

$$\theta(\delta) \equiv E\Big[E[Y|A=1,W(\delta)] - E[Y|A=0,W(\delta)]\Big] = E[Y|A=1,W(\delta)] - E[Y|A=0,W(\delta)]. \tag{17}$$

Under this additional modelling assumption, the targeted maximum-likelihood estimator of $\theta(\delta)$ no longer requires inverse weighting. Specifically, assume that we have available an estimate $g_n(\delta)$ and $Q_n(\delta)$ of the relevant nuisance parameters, with $Q_n(\delta)$ satisfying the additional modelling assumption so that it can be written as $Q_n(\delta) = \beta_n(\delta)A + r_n(W(\delta)) = \theta(\delta) + r_n(W(\delta))$ for some function $r_n(\cdot)$ of $W(\delta)$. As before, the targeted maximum-likelihood estimator is based on an updated estimate $Q_n^1(\delta)$ of $Q(\delta)$ that is obtained by adding a particular covariate $h(\delta)(A, W)$ to that initial fit. In this case, that covariate is given by

$$h(\delta)(A, W) = A - g_n(\delta)(1, W). \tag{18}$$

The updated fit for $Q(\delta)$ can then be written as $Q_n^1(\delta) = [\beta_n(\delta)+\epsilon_n(\delta)]A+r_n^1(W(\delta))$ so that the targeted maximum-likelihood estimator of $\theta(\delta)$ is given by

$$\theta_n^{T-MLE}(\delta) = \beta_n(\delta) + \epsilon_n(\delta). \tag{19}$$

In distinction to the non-parametric targeted maximum-likelihood estimator discussed previously, we will refer to this estimator as the model-based targeted maximum-likelihood estimator. This estimator solves the estimating function

$$0 = \frac{1}{n}\sum_{i=1}^{n} D(\delta)(O_i|g_n(\delta), r_n^1(\delta), \theta(\delta)) \tag{20}$$

corresponding to the estimating function

$$D(\delta)(O|g(\delta), r(\delta), \theta(\delta)) = \Big[A - g(\delta)(1, W)\Big]\Big[Y - \theta(\delta) - r(\delta)(A, W)\Big]. \tag{21}$$

Inference can thus again be based on the influence curve of this estimator. In this case, the standardizing constant c is given by

$$\begin{aligned} c &= -\frac{\partial}{\partial\theta(\delta)} ED(O|g_0(\delta), r_0(\delta), \theta(\delta))\Big|_{\theta(\delta)=\theta_0(\delta)} \\ &= A - g_0(\delta)(1, W). \end{aligned} \tag{22}$$

3 Selection of the targeted adjustment set

While the performance of all four estimators described above can be severely compromised if the ETA assumption is violated (Bembom and van der Laan, 2007b), the problems become most apparent in the case of the IPTW estimator. Unlike the other three estimators that under such circumstances can also rely on extrapolation through a correctly specified model for the regression $Q(\delta)$, the IPTW estimator is based entirely on inverse weighting by an estimate of the treatment mechanism $g(\delta)$, making it thus highly susceptible to violations of the ETA assumption. Under a theoretical violation, a subgroup of the target population is never observed at one of the possible treatment levels, preventing the re-weighting approach from successfully adjusting for confounding and thus resulting in biased estimates. Under a practical violation, observations with very small estimated treatment probabilities $g_n(\delta)$ and corresponding large weights tend to dominate the remainder of the sample so that the estimator becomes highly variable. In addition, it has been shown that a practical violation of the ETA assumption can in fact also cause the IPTW estimator to become biased in finite samples Neugebauer and van der Laan (2005). These considerations suggest that the finite-sample bias of this estimator is a useful measure of the degree to which a departure from the ETA assumption has caused the adjusted variable importance parameter $\theta(\delta)$ to become non-identifiable from the data at hand.

Wang et al. (2006) propose the following simulation-based approach for obtaining an estimate of this bias: The empirical distribution of $W(\delta)$ along with the nuisance parameter estimates $g_n(\delta)$ and $Q_n(\delta)$ define an estimate $P_n(\delta)$ of the data-generating distribution $P(\delta)$ for the observed data structure $O(\delta) = (W(\delta), A, Y)$. Under $P_n(\delta)$, the true value of the adjusted variable importance parameter $\theta(\delta)$ can be obtained by G-computation as

$$\theta(P_n(\delta)) = \frac{1}{n}\sum_{i=1}^{n} Q_n(\delta)(1, W_i) - Q_n(\delta)(0, W_i). \tag{23}$$

At the same time, we can obtain a sampling distribution of IPTW estimates $\theta_{n,1}^{IPTW}(\delta), \ldots, \theta_{n,K}^{IPTW}(\delta)$ by applying the IPTW estimator to a large number K of realizations of the observed data structure $O(\delta)$ that were simulated under $P_n(\delta)$. The finite-sample bias of the IPTW estimator can then be estimated in a straightforward manner by

$$B_{sim}^{ETA}(\delta) = \frac{1}{K}\sum_{k=1}^{K}\theta_{n,k}^{IPTW}(\delta) - \theta(P_n(\delta)). \tag{24}$$

A possible limitation of this parametric bootstrap approach lies in its reliance on a large number of simulated data sets. First, such simulations can be computationally intensive so that the approach would not scale well to applications in which the group of candidate input variables for which we aim to obtain variable importance estimates is large. Second, unless an enormous number of data sets are simulated, the bias estimates can be expected to be quite sensitive to the exact number of simulated data sets used.

A computationally more tractable closed-form measure of finite-sample non-identifiability may be based on the following argument. A common recommendation for increasing the stability of IPTW estimators under practical ETA violations is to truncate the inverse-probability-of-treatment weights $1/g_n(\delta)(A,W)$ by some truncation constant M, thus using weights $wt_M = min(M, 1/g_n(\delta)(A,W))$ instead. Under a practical ETA violation, the use of such truncated weights can lead to a dramatic reduction in the variability of the IPTW estimator, but it typically also increases its bias. As long as at least some of the truncated weights wt_M are strictly less than the original weights $1/g_n(\delta)(A,W)$, the IPTW estimator will in fact often become asymptotically biased. Under a data-generating distribution that satisfies the ETA assumption, however, the estimated treatment probabilities $g_n(\delta)(A,W)$ are clearly bounded away from zero so that M would have to be chosen quite small for the truncated weights to become different from the original weights. Modest levels of truncation corresponding to a reasonably large value of M thus typically do not cause the IPTW estimator to become asymptotically biased if the ETA assumption is satisfied. These considerations suggest that the extent to which the ETA assumption is violated can also be quantified by the asymptotic bias of the IPTW estimator under modest truncation.

Bembom and van der Laan (2007a) recently derived a closed-form estimate for this bias as a function of the truncation constant M. Letting $g_{M,n}(\delta) \equiv max(g_n(\delta), M)$, this estimate is given by

$$\begin{aligned} B_M^{ETA}(\delta) \quad = \quad & \sum_{i=1}^{n} Q_n(\delta)(1,W_i)\frac{g_n(\delta)(1,W_i) - g_{M,n}(\delta)(1,W_i)}{g_{M,n}(\delta)(1,W_i)} - \\ & \sum_{i=1}^{n} Q_n(\delta)(0,W_i)\frac{g_n(\delta)(0,W_i) - g_{M,n}(\delta)(0,W_i)}{g_{M,n}(\delta)(0,W_i)} \end{aligned} \tag{25}$$

While this identifiability criterion is more computationally tractable than the simulation-based finite-sample bias estimate, it also requires the user to supply an appropriate truncation level M. The smaller M is chosen, the more sensitive $B_M^{ETA}(\delta)$ will be to practical ETA violations. At the same time, M should be chosen large enough to ensure that $B_M^{ETA}(\delta) = 0$ under a data-generating distribution that satisfies the ETA assumption. Since the IPTW estimator can tolerate larger weights as sample size increases, it would seem reasonable to make the selected truncation level a function of the sample size. One particular proposal would be to select M such that no observation in the sample would have a weight greater than some proportion p of the sum of all weights, where sensible choices for p corresponding to fairly modest levels of truncation might lie in the range from 0.05 to 0.20. We will examine the sensitivity of this proposed identifiability criterion to the exact choice of M in our data analysis in section 6.

It remains to define a reference with respect to which we evaluate the magnitude of the estimated bias $B_{sim}^{ETA}(\delta)$ or $B_M^{ETA}(\delta)$. A simple choice might be to consider the corresponding point estimate $\theta_n^{T-MLE}(\delta)$. Since the adjusted point estimates can become quite unreliable, however, if the ETA assumption is violated, we suggest to use the unadjusted variable importance estimate $\theta_n^{T-MLE}(0)$ instead. The targeted adjustment set $W^t = W(\delta^t)$ can now be selected by choosing δ^t to be the largest value of δ, $0 \leq \delta \leq 1$, such that the estimated bias is no greater than some proportion B_{max} of the unadjusted variable importance estimate. Here B_{max} is another user-supplied parameter, with reasonable choices likely to be made in the range from 0.10 to 0.50. We note that since the selection of $\theta(\delta^t)$ is made without knowledge of the point estimates $\theta_n^{T-MLE}(\delta)$, inference for $\theta(\delta^t)$ as based on the influence curve remains valid.

4 Selection of the effective adjustment set

Given a targeted adjustment set $W(\delta^t)$, we now aim to select an effective adjustment set $W^e = W(\delta^e)$ such that the corresponding estimator $\theta_n^{T-MLE}(\delta^e)$ has minimal mean squared error for estimating the targeted parameter $\theta(\delta^t)$. In many cases, the effective adjustment set will be equal to the targeted adjustment set, but if the targeted parameter still suffers from a mild practical ETA violation, it is possible that a smaller effective adjustment set will lead to a more efficient estimator.

The mean squared error for an estimator of $\theta(\delta^t)$ can be decomposed into the square of its bias and its variance. Since we will focus here on the non-parametric and model-based targeted maximum-likelihood estimators, the latter component can be estimated in a straightforward way based on the influence of these estimators (see section 2). We will use our estimates of the data-generating distributions $P_n(\delta)$ to obtain an estimate of the bias incurred by using a subset $W(\delta)$ of $W(\delta^t)$ in estimating $\theta(\delta^t)$. Under $P_n(\delta^t)$, the true parameter value of $\theta(\delta^t)$ is given by the G-computation estimate

$$\theta(P_n(\delta^t)) = \frac{1}{n}\sum_{i=1}^{n} Q_n(\delta^t)(1, W_i) - Q_n(\delta^t)(0, W_i) = \theta_n^{G-comp}(\delta^t). \tag{26}$$

Under $P_n(\delta)$, $\delta \leq \delta^t$, the true parameter value of $\theta(\delta)$ is likewise given by

$$\theta(P_n(\delta)) = \frac{1}{n}\sum_{i=1}^{n} Q_n(\delta)(1, W_i) - Q_n(\delta)(0, W_i) = \theta_n^{G-comp}(\delta). \tag{27}$$

The desired bias can thus be estimated by the difference of the two relevant G-computation point estimates:

$$B_n^t(\delta) = \theta_n^{G-comp}(\delta) - \theta_n^{G-comp}(\delta^t) \tag{28}$$

We can now select δ^e as the minimizer over $0 \leq \delta \leq \delta^t$ of the corresponding mean squared error estimate

$$MSE_n^t(\delta) = \left[B_n^t(\delta)\right]^2 + V_n(\delta), \tag{29}$$

where $V_n(\delta)$ is an estimate of the variance of $\theta_n^{T-MLE}(\delta)$ as based on the influence curve of that estimator.

Since the selection of $\theta(\delta^e)$ is based on knowledge of the point estimates $\theta_n^{G-comp}(\delta)$, honest inference for the resulting estimator would have to take into account that it was selected from among several candidate estimators, specifically with the goal of minimizing mean squared error. Inference based on the influence curve of $\theta_n^{T-MLE}(\delta^e)$ may thus be somewhat optimistic since it ignores the data-adaptive selection of the estimator. Honest inference could be obtained based on a bootstrap procedure that includes this estimator selection step. Since we have that $B_n^t(\delta^t) = 0$, we note, however, that $\theta(\delta^e)$ can be expected to converge to $\theta(\delta^t)$ so that inference based on the influence curve of $\theta_n^{T-MLE}(\delta^e)$ remains asymptotically valid.

5 Simulation study

The selection of the targeted adjustment set is a question of selecting the scientific parameter of interest. The practical performance of the proposed approach to this problem is therefore better illustrated in an applied data analysis than in a simulation study. In this section, we present a simulation study that is aimed at examining to what extent the performance of the non-parametric and model-based targeted maximum-likelihood estimators of a given targeted variable importance parameter $\theta(\delta^t)$ can be improved by the data-adaptive selection of an effective adjustment set $W(\delta^e)$.

For this purpose, we consider a point-treatment data structure $O = (W, A, Y)$, with $W = (W_1, \ldots, W_{10})$ containing ten potential confounding factors, A denoting a binary treatment variable, and Y representing a continuous outcome of interest. Given a treatment mechanism $g_0(A \mid W)$ and the regression function $Q_0(A, W)$, the observed data structure was generated as follows:

1. Generate $W_1, \ldots, W_{10}$ as independent random uniform variables over the interval $[0, 1]$.
2. Generate the observed treatment variable A from the conditional distribution of A given W, $g_0(A \mid W)$.
3. Generate the observed outcome Y as $Y = Q_0(A, W) + \epsilon$ with $\epsilon \sim N(0, 1)$.

We consider the two different treatment mechanism

$$\text{logit}\Big(g_{1,0}(A \mid W)\Big) = W_3 - W_4 + 2W_5 - 2W_6 + 2W_7 - 2W_8 - 3W_9 + 4W_{10} \tag{30}$$

and

$$\text{logit}\Big(g_{2,0}(A \mid W)\Big) = W_3 - W_4 + 2W_5 - 2W_6 + 2W_7 - 2W_8 - 2W_9 + 2W_{10}. \tag{31}$$

The regression function $Q_0(A, W)$ is given by

$$Q_0(A, W) = A + 2W_2 + 2W_3 + 2W_4 + W_7 + W_8 + 0.1W_9 + 0.1W_{10}. \tag{32}$$

We thus have two different data-generating distributions $(g_{1,0}, Q_0)$ and $(g_{2,0}, Q_0)$. The targeted parameter is given by the fully adjusted marginal variable importance

$$\theta = E\Big[E[Y|A = 1, W] - E[Y|A = 0, W]\Big]. \tag{33}$$

Under $(g_{1,0}, Q_0)$, the covariates W_9 and W_{10} are strong predictors of A so that they may cause a moderate practical violation of the ETA assumption. Since they have only a weak effect on Y, omitting these two covariates from the effective adjustment set might therefore lead to a considerable reduction in the variability of the estimator, at the price of only a slight increase in bias. Data-adaptive selection of an effective adjustment set can therefore be hoped to lead to a significant increase in efficiency under this data-generating distribution. Under $(g_{2,0}, Q_0)$, W_9 and W_{10} are only moderate predictors of A so that much smaller efficiency gains might be expected under this data-generating distribution.

Table 1: Mean squared error of the non-parametric and model-based targeted maximum-likelihood estimators using either the targeted adjustment set or a data-adaptively selected effective adjustment set.

	Non-parametric		**Model-based**	
	Targeted	Effective	Targeted	Effective
$(g_{1,0}, Q_0)$				
n = 100	16.7632	0.0917	0.0758	0.0670
n = 500	0.0312	0.0140	0.0138	0.0131
n = 2500	0.0051	0.0025	0.0027	0.0025
$(g_{2,0}, Q_0)$				
n = 100	1.9750	0.0828	0.0645	0.0621
n = 500	0.0168	0.0125	0.0119	0.0116
n = 2500	0.0030	0.0022	0.0022	0.0021

Table 1 summarizes the mean-squared errors for the non-parametric and model-based targeted maximum-likelihood estimators of θ using either the targeted adjustment set or a data-adaptively selected effective adjustment set for three different sample sizes. As expected, the fully adjusted non-parametric estimator is more sensitive to practical ETA violations than the fully adjusted model-based estimator, with its variance being considerably greater under $(g_{1,0}, Q_0)$ than under $(g_{2,0}, Q_0)$. Consequently, the non-parametric estimator also benefits much more strongly from the data-adaptive selection of an effective adjustment set, with efficiency gains relative to the fully adjusted estimator of roughly 100% under $(g_{1,0}, Q_0)$ and 35% under $(g_{2,0}, Q_0)$ for sample sizes of $n = 500$ and greater. The enormous efficiency gains observed for this estimator for $n = 100$ suggest a considerable practical ETA violation that in practice might have resulted in the selection of a smaller targeted adjustment set. The efficiency gains for the model-based estimators are slight compared to those for the non-parametric estimator. Since the assumption of no interaction between A and W is satisfied in this simulation study, the model-based estimator is typically more efficient than the non-parametric estimator. We note, however, that the performance of the two estimators based on a data-adaptively selected effective adjustment set is comparable, especially as sample size increases, which is another testament to the considerable efficiency gains achieved by the non-parametric estimator.

6 Data analysis

In this section we apply the methodology described above to the task of identifying mutations in the protease enzyme of HIV that modulate how well the virus can replicate in the presence of a particular antiretroviral drugs, and thus how well a patient responds to that drug. A considerable number of such drugs are available for treating patients infected with HIV, with the main mechanistic classes consisting of protease inhibitors (PIs), nucleotide and nucleoside reverse transcriptase inhibitors (NRTIs), and nonnucleoside reverse transcriptase inhibitors (NNRTIs). While a patient is being treated with a particular combination of these drugs, the virus frequently acquires a number of mutations that reduce its susceptibility to that drug regimen, requiring the patient to be switched to a new regimen that the virus remains sensitive to. When faced with this situation, clinicians frequently genotype the virus to ascertain the presence or absence of a large number of mutations that are thought to contribute to the resistance to various drugs (Shafer et al., 2000). This practice motivates us here to identify in a systematic way mutations that have a strong impact on a patient's virologic response to a new drug treatment and that could thus guide a clinician in designing a salvage therapy regimen on the basis of genotypic test results.

The effect of viral mutations on virologic response to therapy can be seriously confounded by a patient's treatment history. Past treatment regimens exert a strong selection pressure on viral evolution, thus affecting the probability that a given mutation is observed. In addition, treatment history can have an independent impact on virologic response by resulting in archived, or latent, virus carrying unobserved mutations that affect response to subsequent treatment regimens. As a result, an unadjusted association observed between a given mutation and treatment response may in fact be due to the presence of other mutations, both observed and unobserved. Treatment strategies vary across populations and evolve over time, potentially resulting in distinct mutation distributions. Thus, control of confounding due to treatment history is needed to ensure that the estimated importance of a given mutation can be more readily generalized to populations other than the original study population.

Similarly, we would also like to adjust for the presence of additional mutations. Mutations conferring resistance to drugs of a class different from that targeted by the mutation of interest, thus affecting a distinct viral enzyme, can typically be controlled for without much difficulty. However, mutations conferring resistance to the same drug class, thus affecting the same viral enzyme, are often so strongly correlated that the corresponding adjusted variable importance parameter is subject to a severe ETA violation. This is due to the fact that, while correlation between mutations affecting distinct viral enzymes occurs primarily as a result of past treatment patterns, correlation between mutations in the same enzyme often occurs as part of an evolutionary pathway towards resistance to drugs targeting that enzyme. Previous analyses have typically addressed this problem by categorically not adjusting for any mutations affecting the same viral enzyme as the mutation under consideration (Bembom et al., 2007). One might expect, however, that only a subset of the mutations affecting the same viral enzyme are so strongly correlated with the mutation under consideration as to cause serious ETA problems so that data-adaptive selection of the adjustment set might lead to variable importance estimates that typically suffer from less confounding than those obtained in earlier analyses.

6.1 Data source

Analyses were based on a data set, described previously in Bembom et al. (2007), consisting of observational clinical data that were primarily drawn from the Stanford drug resistance database and supplemented with data from an ongoing collaboration with the Kaiser Permanente Medical Care Program, Northern California. Currently, the Stanford database contains longitudinal data on over 6,000 patients. Data collected include use of antiretroviral drugs, results of viral genotype tests, and measurements of viral load as well as CD4 T cell count collected during the course of clinical care.

For the sake of illustration, we focus on resistance to the commonly used PI drug lopinavir. We identified all Treatment Change Episodes (TCEs) in this database that involved initiation of a salvage regimen containing lopinavir. A TCE was defined using the following inclusion criteria: 1) change of at least one drug from the patient's previous antiretroviral regimen; 2) availability of a baseline viral load and genotype within 24 weeks prior to the change in regimen; and, 3) availability of an outcome viral load 4-36 weeks after the change in regimen and prior to any subsequent changes in regimen.

TCEs were excluded if no candidate resistance mutations were present in the baseline genotype, if the subject had no past experience of PI drugs prior to the current regimen, or if the newly initiated regimen

included hydroxyurea, any experimental antiretroviral drugs, or any PI drugs other than lopinavir (apart from the low dose of ritonavir that is always given with lopinavir). If a single baseline genotype had several subsequent regimen changes that met inclusion criteria as TCEs, only the first of these regimen changes was included in analyses. Multiple TCEs, each corresponding to a unique baseline genotype, treatment changes, and outcome, were allowed from a single individual; the resulting dependence between TCEs was accounted for in the derivation of standard errors and p-values. Based on these inclusion criteria, we identified 401 TCEs among 372 subjects that were included in our analyses. We considered as candidate biomarkers all mutations assessed by the Stanford HIVdb algorithm to be potentially related to resistance to any approved PI drug (`http://hivdb.stanford.edu`, accessed 9/1/2007). Including only mutations that occurred in at least two TCEs, we are faced with a total of 26 candidate PI mutations.

Antiretroviral regimens generally combine drugs from more than one class. The following characteristics of the non-PI component of the salvage regimen were therefore included in the set W of potential adjustment variables: indicators of use of each of 13 non-PI drugs; number of drugs used in each major non-PI class; number of drugs and number of classes used in the salvage regimen for the first time; use of an NNRTI drug in the salvage regimen for the first time; and number of drugs switched between the previous and salvage regimen. W also included the following covariates collected prior to the baseline genotype: indicators of past treatment with each of 30 antiretroviral drugs; number of drugs used in each of the three major drug classes (PI, NRTI, and NNRTI); history of mono or dual therapy; number of past drug regimens; date of earliest antiretroviral therapy; highest prior viral load; lowest prior CD4 T cell count; and most recent (baseline) viral load.

The covariate set W also included indicators for the presence or absence of PI mutations other than the mutation of interest itself as well as indicators for the presence or absence of known NRTI and NNRTI mutations. In addition, we included summaries of the non-PI mutations in the baseline genotype. Known NRTI and NNRTI resistance mutations present at baseline were summed. Furthermore, susceptibility scores (standardized to a 0-1 scale) were calculated for each non-PI antiretroviral drug using the Stanford HIVdb scoring system. These susceptibility scores were included both as individual covariates and as interactions with indicators of the use of their corresponding drugs in the salvage regimen. Finally, these interaction terms were summed to yield a non-PI genotypic susceptibility score (GSS), which summarized the activity of the non-PI component of the regimen. The set of potential adjustment variables W included a total of 163 variables.

The outcome of interest, clinical virologic response, could be conceived as either a binary indicator of success (defined as achievement of a final viral load below the assay's lower limit of detection of 50 copies/mL), or as a continuous measure such as the change in final log_{10} viral load over baseline log_{10} viral load. The analyses reported here used a hybrid of these two approaches, aiming to capture the strengths of each. Specifically, given a baseline measurement Y_0 and a follow-up measurement Y_1 of log_{10} viral load, the outcome of interest Y was defined as follows: If Y_1 was above the lower limit of detection ($Y_1 > 1.7$), then $Y = Y_1 - Y_0$; if Y_1 was below the detectability limit, however, we imputed Y as the maximum decrease in viral load detected in the population, which was -4.2 log. Under this definition, both large drops in viral load from a high baseline and any achievement of an undetectable viral load (regardless of baseline) were treated as clinical successes. When several viral loads were measured between 4 and 36 weeks after regimen change, the first was used; duration from initiation of the salvage regimen until outcome measurement was included in the adjustment set W.

6.2 Variable importance estimation

The goal of our analysis was to estimate the impact of each of the 26 candidate PI mutations on Y, adjusting for as many elements of W as possible, and to rank the mutations based the statistical evidence for a non-zero variable importance. For this purpose we focus on the non-parametric and model-based targeted maximum-likelihood estimators described in section 2. We compare the results based on data-adaptively selected targeted and effective adjustment sets to those based on unadjusted and fully adjusted analyses.

Covariates that are not predictive of the outcome of interest neither confound the effect of a mutation on viral nor have the potential to increase the precision with which that effect can be estimated. Hence we first carried out a dimension reduction step aimed at identifying those covariates in W that appear to be associated with viral load. For this purpose, we examined the univariate association between each baseline covariate W_j and Y using a univariate repeated-measures regression. In this manner, we

obtained p-values for the null hypotheses that a given covariate is independent of Y. Since the collection of candidate baseline covariates was sizeable, these marginal p-values were adjusted for the simultaneous performance of multiple hypothesis tests using the approach developed by Benjamini and Hochberg (1995) for controlling the false discovery rate (FDR). Out of the 163 variables contained in W, we retained a total of 51 that remained significantly associated with Y at a significance level of 0.05.

Following this dimension reduction, we applied the Deletion/Substitution/Addition (D/S/A) algorithm (Sinisi and van der Laan, 2004) to obtain estimates of the two nuisance parameters $g(\delta)$ and $Q(\delta)$. The D/S/A algorithm is a data-adaptive algorithm for polynomial regression that generates candidate predictors as linear combinations of polynomial tensor products in the candidate explanatory variables. These candidate estimators are indexed by the number and complexity of the terms, and the optimal candidate is selected using cross-validation. A version of the D/S/A algorithm was employed that relied solely on addition moves to generate candidate estimators, thus making it similar to a forward regression approach except that the size of the estimator is selected by cross-validation rather than by p-values; deletion and substitution moves were omitted to reduce computational complexity. The algorithm considered candidate estimators consisting of up to 20 terms.

Given a set of candidate explanatory variables, two-way interactions were explored based on repeated-measures regression models aimed at predicting Y as function of two candidate explanatory variables as well as the corresponding interaction term. Two-way interaction terms that were significant at an FDR-adjusted significance level of 0.05 were explicitly included in the set of candidate explanatory variables. The D/S/A algorithm was then allowed to consider estimators consisting only of main-effect terms taken from that set of candidate explanatory variables. This approach of not considering candidate estimators involving arbitrary two-way interactions is motivated not only by computational considerations, but also by the observation that such estimators are typically far more variable than those based on main-effect terms only, thus often leading to the selection of estimators including only main-effect terms. Including important interaction terms explicitly in the set of explanatory variables can thus be hoped to alleviate the discontinuity in variability seen in moving from estimators consisting of only main-effect terms to those involving also two-way interactions, thus increasing the chance that important two-way interaction terms will be selected in the final estimator.

Since we are interested in estimating the effect of a given mutation A on Y, we would like A to be included in the regression model for $E[Y|A,W]$. Forcing A into the model and then allowing the D/S/A algorithm to add elements of W is problematic, however, since adjustment variables that are strongly correlated with A may contribute little to the accurate prediction of Y once A is included in the model. Such an approach might thus lead to important confounding factors being omitted from the model. We therefore first allow the D/S/A algorithm to data-adaptively select a linear regression model for $E[Y|W]$ before then re-fitting that model with A added to the selected explanatory variables.

The D/S/A algorithm was also used to select an appropriate logistic regression model for the treatment mechanism $g(\delta)$. The selection criterion of minimizing cross-validated risk employed by the D/S/A algorithm for selecting the size of the estimator is aimed at selecting an estimator with good prediction properties. Optimizing the bias-variance trade-off for this purpose often leads to estimators consisting of only a small number of terms, causing the selected regression fit for $g(\delta)$ to give an unrealistically optimistic impression of the extent to which the ETA assumption is satisfied. For this reason, it is typically advisable to use somewhat more non-parametric estimates of the treatment mechanism for the task of assessing the validity of the ETA assumption (Wang et al., 2006). We therefore selected the size of the regression fit for g not as the minimizer of cross-validated risk, but rather as the largest size such that the corresponding cross-validated risk was no more than 25% above the minimal cross-validated risk. Theoretical arguments in fact imply that such slight overfits of the treatment mechanism will in first order also increase the efficiency of the resulting variable importance estimator (van der Laan and Robins, 2003). Overfits may in theory negatively affect the performance of the estimator through second-order terms if some of the estimated treatment probabilities become very close to zero, but the variable importance algorithm proposed here addresses that problem by selecting a targeted adjustment set for which the ETA assumption is well approximated. In addition, we set estimated treatment probabilities smaller than 0.01 to 0.01.

Point estimates based on the approach presented in section 2 for a sample of n i.i.d. observations remain valid in the context of repeated measures. The efficiency of the estimator might be improved by optimizing the weights given to individual observations on basis of an estimated correlation matrix for observations obtained from the same subject, as is done in generalized estimating equations Liang

and Zeger (1986), but given the small number of repeated measures in the data set at hand we simply give equal weights to all observations. Estimation is thus based on estimating functions that are a sum over all the observations contributed by a single subject. Assuming that the number of observations contributed by each subject is non-informative, inference can be based in a straightforward manner on these modified estimating functions.

6.3 Unadjusted and fully adjusted variable importance estimates

Among the 26 candidate PI mutations considered here, the Stanford scoring system identifies the following 12 mutations as major contributors to lopinavir resistance: 50V, 82AFST, 46ILV, 54VA, 54LMST, 84AV, 32I, 47V, 48VM, 82MLC, 84C, and 90M; the remaining 14 mutations are thought to make minor or no contributions to resistance. Here and subsequently, mutations are denoted by the position of the change in the HIV protease enzyme, followed by a letter indicating the amino acid that has been substituted (e.g. 53LY refers to a substitution of leucine or tyrosine at protease position 53).

The unadjusted variable importance analysis, summarized in table 2, yielded significant p-values for all but eight of the candidate PI resistance mutations. Four of these eight mutations occurred in fewer than 10 TCEs so that the analysis had low power to detect an impact of these mutations on viral load. Among these four mutations were two mutations, 82MLC and 84C, that are thought to have a major effect on lopinavir resistance. The significant subset includes the remaining 10 known major lopinavir resistance mutations, but also eight mutations thought to make minor or no contributions to resistance. Among these were the mutations 30N, 88DTG, and 88S, all estimated to be significantly protective. Under the Stanford scoring system, mutations only receive a score if they are thought to increase resistance to a particular drug so that these findings are not necessarily in disagreement with the scores of zero assigned to these three mutations by that system. It seems quite plausible that mutations may also decrease the fitness of the virus and thus lead to improved virologic response. In fact, *in vitro* experiments examining the effect of different mutations on viral phenotype suggest that 30N and 88S may in fact have a negative impact on the fitness of the virus (Rhee et al., 2006). The significant subset still contains five mutations, however, that are estimated to be associated with considerably worse virologic response, but are not considered major lopinavir drug resistance mutations by the Stanford scoring system (33F, 73CSTA, 10FIRVY, 20IMRTVL, and 71ITV). Two of these, 33F and 73CSTA, are in fact ranked among the five most important mutations by the unadjusted analysis, illustrating that an analysis not addressing the issue of confounding can lead to rather noisy results.

An analysis based on the non-parametric targeted maximum-likelihood estimator adjusting for the full set W of potential confounders, summarized in table 3, identifies only five mutations as having a major effect on virologic response to lopinavir (50V, 84C, 16E, 32I, and 48VM). In agreement with the Stanford scoring system, the two mutations 50V and 32I are estimated to lead to decreased susceptibility to lopinavir. The mutation 16E, estimated to lead to considerably improved fitness of the virus in the presence of lopinavir, is not thought to be a major contributor to lopinavir drug resistance. The remaining two mutations 84C and 48VM, finally, are thought to be major contributors, but are in fact estimated to lead to increased susceptibility to lopinavir. The ranking produced by this analysis is thus hard to reconcile with current understanding of HIV antiretroviral resistance, illustrating that a fully adjusted analysis can lead to unreliable results if the ETA assumption is violated. A fully adjusted analysis based on the model-based targeted maximum-likelihood estimator, summarized in table 4, identified no mutations with a statistically significant impact on virologic response. These findings are similarly unsatisfying and show that a violation of the ETA assumption cannot be adequately addressed by turning to a more stable estimator that is based on additional modelling assumptions.

6.4 Data-adaptive selection of the targeted and effective adjustment sets

In this section, we examine the variable importance estimates obtained by data-adaptive selection of the targeted and effective adjustment set. In section 3, we proposed two different criteria for selecting the targeted adjustment set, one based on a simulation aimed at estimating the finite-sample bias of the IPTW estimator, the other based on a closed-form estimate of the asymptotic bias of a modestly truncated IPTW estimator. The latter criterion depends on a user-supplied choice for the parameter p that defines the desired truncation level based on the maximum proportion of the sum of all weights

Table 2: Unadjusted estimates variable importance estimates ranked by p-value. *The table also shows the resistance score assigned to a mutation by the Stanford HIVdb scoring system (accessed on 9/1/2007) and the frequency of the mutation among the 401 treatment change episodes.*

Rank	Mutation	Score	Freq	Estimate	SE	p-value
1	30N	0	45	-1.12	0.23	0.0000
2	54VA	11	84	0.86	0.18	0.0000
3	50V	20	5	1.98	0.44	0.0001
4	33F	5	44	0.84	0.21	0.0005
5	73CSTA	2	66	0.81	0.22	0.0012
6	88DTG	0	44	-0.89	0.25	0.0012
7	82AFST	20	100	0.62	0.18	0.0016
8	10FIRVY	2	217	0.54	0.16	0.0022
9	90M	10	171	0.54	0.16	0.0028
10	47V	10	17	1.18	0.36	0.0029
11	54LMST	11	36	0.69	0.24	0.0110
12	88S	0	9	-0.74	0.27	0.0134
13	32I	10	21	0.80	0.30	0.0146
14	20IMRTVL	0	115	0.46	0.18	0.0182
15	46ILV	11	143	0.43	0.17	0.0182
16	84AV	11	73	0.49	0.20	0.0219
17	71TVI	2	181	0.36	0.16	0.0312
18	48VM	10	16	0.77	0.35	0.0378
19	53LY	3	33	0.53	0.26	0.0601
20	24IF	2	16	0.69	0.36	0.0691
21	36ILVTA	0	141	0.32	0.18	0.0919
22	23I	0	4	0.68	1.02	0.5950
23	63P	0	311	0.09	0.19	0.7297
24	82MLC	10	4	0.30	0.95	0.8123
25	16E	0	9	-0.05	0.50	0.9308
26	84C	10	2	0.15	1.74	0.9308

Table 3: Fully adjusted non-parametric variable importance estimates ranked by p-value. *The table also shows the resistance score assigned to a mutation by the Stanford HIVdb scoring system (accessed on 9/1/2007) and the frequency of the mutation among the 401 treatment change episodes.*

Rank	Mutation	Score	Freq	Estimate	SE	p-value
1	50V	20	5	0.95	0.08	0.0000
2	84C	10	2	-2.92	0.08	0.0000
3	16E	0	9	0.90	0.25	0.0021
4	32I	10	21	0.26	0.07	0.0021
5	48VM	10	16	-0.26	0.07	0.0021
6	88S	0	9	-0.35	0.15	0.0708
7	33F	5	44	1.17	0.51	0.0823
8	54VA	11	84	0.55	0.28	0.1583
9	84AV	11	73	0.44	0.28	0.3457
10	53LY	3	33	0.51	0.38	0.4573
11	30N	0	45	-0.22	0.18	0.5307
12	47V	10	17	0.76	0.65	0.5307
13	10FIRVY	2	217	-0.22	0.21	0.6000
14	23I	0	4	-0.86	1.01	0.7144
15	73CSTA	2	66	0.20	0.25	0.7144
16	82AFST	20	100	-0.25	0.36	0.7824
17	82MLC	10	4	-0.33	0.61	0.9031
18	20IMRTVL	0	115	0.08	0.19	0.9098
19	36ILVTA	0	141	0.09	0.22	0.9098
20	90M	10	171	0.20	0.46	0.9098
21	63P	0	311	-0.09	0.29	0.9120
22	88DTG	0	44	-0.17	0.53	0.9120
23	46ILV	11	143	-0.04	0.21	0.9371
24	54LMST	11	36	-0.04	0.22	0.9371
25	71TVI	2	181	0.02	0.18	0.9651
26	24IF	2	16	0.00	0.27	0.9967

Table 4: Fully adjusted model-based variable importance estimates ranked by p-value. *The table also shows the resistance score assigned to a mutation by the Stanford HIVdb scoring system (accessed on 9/1/2007) and the frequency of the mutation among the 401 treatment change episodes.*

Rank	Mutation	Score	Freq	Estimate	SE	p-value
1	50V	20	5	1.35	0.54	0.2621
2	54VA	11	84	0.57	0.25	0.2621
3	16E	0	9	0.45	0.36	0.6008
4	24IF	2	16	0.54	0.31	0.6008
5	30N	0	45	-0.38	0.31	0.6008
6	33F	5	44	0.36	0.31	0.6008
7	36ILVTA	0	141	0.24	0.19	0.6008
8	47V	10	17	0.57	0.52	0.6008
9	48VM	10	16	-0.44	0.36	0.6008
10	53LY	3	33	0.30	0.26	0.6008
11	73CSTA	2	66	0.39	0.25	0.6008
12	88S	0	9	-0.47	0.33	0.6008
13	32I	10	21	0.37	0.36	0.6102
14	10FIRVY	2	217	-0.16	0.18	0.7016
15	20IMRTVL	0	115	0.08	0.16	0.8481
16	54LMST	11	36	0.17	0.30	0.8481
17	63P	0	311	-0.07	0.18	0.8481
18	71TVI	2	181	0.06	0.15	0.8481
19	84AV	11	73	0.11	0.22	0.8481
20	84C	10	2	-0.38	0.80	0.8481
21	88DTG	0	44	0.11	0.30	0.8481
22	90M	10	171	0.13	0.19	0.8481
23	23I	0	4	-0.23	1.22	0.9214
24	82MLC	10	4	-0.18	0.79	0.9214
25	46ILV	11	143	0.02	0.18	0.9275
26	82AFST	20	100	0.03	0.26	0.9275

that any one weight is allowed to reach. We first examine the sensitivity of the proposed algorithm to different choices for selecting the targeted adjustment set. Table 5 summarizes the targeted adjustment level δ^t selected by the simulation-based criterion as well as by the closed-form criterion for three different choices of p. Overall, the choices made by the four different approaches are in good agreement with each other, with major discrepancies observed only for the mutation 24IF. As is to be expected, larger choices for p, corresponding to milder truncation levels, decrease the sensitivity of the closed-form criterion and thus tend to lead to slightly larger targeted adjustment levels, although the effect is not too strong over the range of candidate values for p considered here.

Table 5: The targeted adjustment level δ^t selected based on the simulation-based criterion $B^{ETA}_{sim}(\delta)$ as well as based on the asymptotic criterion $B^{ETA}_{M}(\delta)$ for $p = 0.05$, $p = 0.10$, and $p = 0.20$. *The maximally tolerated proportion of bias relative to the unadjusted estimate, B_{max}, is set to 0.25.*

Mutation	Simulation	$p = 0.05$	$p = 0.10$	$p = 0.20$
10FIRVY	1.0	1.0	1.0	1.0
16E	0.0	0.0	0.0	0.0
20IMRTVL	1.0	1.0	1.0	1.0
23I	0.0	0.3	0.3	0.3
24IF	0.3	0.1	0.7	0.7
30N	0.5	0.5	0.5	0.5
32I	0.5	0.5	0.5	0.5
33F	0.8	0.7	0.8	0.9
36ILVTA	1.0	1.0	1.0	1.0
46ILV	1.0	1.0	1.0	1.0
47V	0.5	0.5	0.7	0.7
48VM	0.5	0.5	0.5	0.5
50V	1.0	1.0	1.0	1.0
53LY	0.9	0.8	0.9	0.9
54LMST	0.3	0.3	0.3	0.5
54VA	1.0	0.8	0.9	0.9
63P	0.4	0.6	0.6	0.6
71TVI	1.0	1.0	1.0	1.0
73CSTA	0.7	0.6	0.7	0.7
82AFST	0.8	0.7	0.7	0.8
82MLC	0.0	0.2	0.2	0.2
84AV	0.5	0.5	0.8	0.8
84C	0.0	0.0	0.0	0.0
88DTG	0.8	0.7	0.7	0.8
88S	0.0	0.2	0.4	0.4
90M	1.0	1.0	1.0	1.0

Table 6 summarizes the mutations that are identified as having a significant impact on virologic response if the targeted adjustment level is selected based on the same four different approaches. Again, the results are good agreement with each other, especially for those mutations that are thought to have major effect on virologic response. These findings, together with the results displayed in table 5, thus suggest that the proposed algorithm is fairly robust with respect to choices made at this step.

Table 7 summarizes the number of mutations that are statistically significant at the 0.05 level if the effective adjustment is either set equal to the targeted adjustment or selected data-adaptively based on the mean-squared-error criterion described in section 4. As seen already in the simulation study, the gains achieved by the non-parametric estimator are considerably larger than those achieved by the model-based estimator. The non-parametric estimator becomes more sensitive to the approach taken for selecting the targeted adjustment set if the effective adjustment set is not selected data-adaptively, with the number

Table 6: Mutations that have a statistically significant impact on viral load at the 0.05 significance level. *Significant mutations are shown by check marks for the non-parametric (NP) as well as model-based (MOD) estimator and targeted adjustment levels δ^t selected based on the simulation-based criterion $B_{sim}^{ETA}(\delta)$ as well as based on the asymptotic criterion $B_M^{ETA}(\delta)$ for $p = 0.05$, $p = 0.10$, and $p = 0.20$. The maximally tolerated proportion of bias relative to the unadjusted estimate, B_{max}, is set to 0.25. The effective adjustment set is selected data-adaptively. The table also shows the resistance score assigned to a mutation by the Stanford HIVdb scoring system (accessed on 9/1/2007) and the frequency of the mutation among the 401 treatment change episodes.*

Mutation	Score	Freq	Simulation	$p = 0.05$	$p = 0.10$	$p = 0.20$
10FIRVY	2	217				
16E	0	9				
20IMRTVL	0	115				
23I	0	4				
24IF	2	16	✓		✓	✓
30N	0	45	✓	✓	✓	✓
32I	10	21	✓	✓	✓	✓
33F	5	44				
36ILVTA	0	141				
46ILV	11	143				
47V	10	17	✓	✓	✓	✓
48VM	10	16	✓	✓	✓	✓
50V	20	5	✓	✓	✓	✓
53LY	3	33				
54LMST	11	36	✓	✓	✓	✓
54VA	11	84	✓	✓	✓	✓
63P	0	311				
71TVI	2	181				
73CSTA	2	66				
82AFST	20	100	✓	✓	✓	✓
82MLC	10	4				
84AV	11	73	✓	✓	✓	✓
84C	10	2				
88DTG	0	44				
88S	0	9	✓	✓	✓	✓
90M	10	171				

of significant mutations identified by that estimator ranging from three to eight depending on the choice of the identifiability criterion. The model-based estimator, on the other hand, remains relatively stable with respect to that choice even if the effective adjustment set is not selected data-adaptively.

Table 7: Number of mutations that are statistically significant at the 0.05 significance level if the effective adjustment set is selected data-adaptively versus being set equal to the targeted adjustment set. *Results are shown for the non-parametric and model-based estimator as well as for the targeted adjustment set selected based on the simulation-based criterion* $B_{sim}^{ETA}(\delta)$ *as well as based on the asymptotic criterion* $B_{M}^{ETA}(\delta)$ *for* $p = 0.05$*,* $p = 0.10$*, and* $p = 0.20$*. The maximally tolerated proportion of bias relative to the unadjusted estimate,* B_{max}*, is set to 0.25.*

	Non-parametric		**Model-based**	
	Targeted	Effective	Targeted	Effective
Simulation	8	11	13	13
$p = 0.05$	8	10	11	12
$p = 0.10$	4	11	11	13
$p = 0.20$	3	11	10	11

Tables 8 and 9 summarize the variable importance estimates obtained by the algorithm proposed here, selecting the targeted adjustment by the closed-form criterion with $p = 0.05$. The non-parametric estimator identifies 10 mutations with a statistically significant impact on viral load. With the exception of 32I and 88S, all of these 10 mutations are also significant if the effective adjustment set is not selected data-adaptively. Among the 10 identified mutations are eight of the 12 major known drug resistance mutations for lopinavir (50V, 48VM, 47V, 54LMST, 32I, 54VA, 84AV, and 82AFST) as well as two mutations that are estimated to increase susceptibility to lopinavir (30N and 88S), a finding that, as mentioned earlier, is in agreement with *in vitro* experiments examining the effect of different mutations on viral phenotype (Rhee et al., 2006). The same experiments suggest that the mutations 46ILV and 90M, two of the four major mutations not identified by this analysis, may in fact be less important for lopinavir resistance than previously thought. The remaining two important mutations not identified here, 82MLC and 84C, occurred among only four and two of the 401 TCEs used in this analysis, respectively, so that the analysis had very low power for finding a significant impact of these mutations on viral load. Overall, the results reported here are thus in excellent agreement with current understanding of HIV antiretroviral resistance.

The variable importance estimates obtained by the model-based estimator are overall very similar to those obtained by the non-parametric estimator. The significant subset is identical except that the major mutation 84AV is missing and the three minor mutations 33F, 73CSTA, and 88DTG are included. With the exception of 88S, all of the identified 12 mutations are also significant if the effective adjustment set is not selected data-adaptively. The mutation 88DTG is estimated to increase susceptibility to lopinavir so that inclusion of this mutation is not necessarily in disagreement with the Stanford scoring system. The remaining three differences between the significant subset identified here and that described for the non-parametric estimator, however, cause the results for the model-based estimator to be in not quite as strong an agreement with current knowledge about lopinavir drug resistance as those for the non-parametric estimator.

For each of the mutations identified by the non-parametric estimator as a having a significant impact on viral load, table 10 summarizes which of the other significant mutations could not be adjusted for in obtaining an adjusted variable importance estimate. The table illustrates that adjustment for all other mutations is in fact difficult in most cases. Individual contributions to drug resistance are particularly hard to disentangle since mutations thought to decrease sensitivity to lopinavir are typically positively correlated with each other, but negatively with those mutations thought to increase sensitivity. For most candidate PI mutations it is still possible, however, to adjust for the majority of the other mutations. This may explain why the results reported here are in better agreement with current understanding of lopinavir resistance than those reported in previous analyses that categorically did not adjust for any other candidate PI mutations (Bembom et al., 2007). It seems somewhat surprising that even mutations with relatively small marginal correlations with the mutation of interest could sometimes not be adjusted

Table 8: Data-adaptively adjusted non-parametric variable importance estimates ranked by p-value. *The targeted adjustment level is selected based on the asymptotic bias estimate $B_M^{ETA}(\delta)$ for a truncated IPTW estimator. The parameters p and B_{max} are set to 0.05 and 0.25, respectively. The effective adjustment set is selected data-adaptively. δ^t and δ^e give the proportion of potential confounders contained in the targeted and effective adjustment set, respectively. The table also shows the resistance score assigned to a mutation by the Stanford HIVdb scoring system (accessed on 9/1/2007) and the frequency of the mutation among the 401 treatment change episodes.*

Rank	Mutation	Score	Freq	Estimate	SE	δ^t	δ^e	p-value
1	50V	20	5	0.95	0.08	1.0	1.0	0.0000
2	48VM	10	16	1.20	0.18	0.5	0.5	0.0000
3	47V	10	17	1.62	0.25	0.5	0.4	0.0000
4	30N	0	45	-1.12	0.22	0.5	0.5	0.0000
5	54LMST	11	36	0.60	0.17	0.3	0.3	0.0025
6	32I	10	21	0.89	0.26	0.5	0.3	0.0027
7	88S	0	9	-0.74	0.27	0.2	0.0	0.0230
8	54VA	11	84	0.43	0.16	0.8	0.8	0.0251
9	84AV	11	73	0.44	0.17	0.5	0.5	0.0251
10	82AFST	20	100	0.38	0.15	0.7	0.6	0.0389
11	53LY	3	33	0.56	0.25	0.8	0.3	0.0521
12	73CSTA	2	66	0.54	0.24	0.6	0.5	0.0521
13	24IF	2	16	0.64	0.32	0.1	0.1	0.0947
14	33F	5	44	0.55	0.29	0.7	0.7	0.1068
15	36ILVTA	0	141	0.27	0.15	1.0	0.5	0.1333
16	90M	10	171	0.30	0.17	1.0	0.8	0.1333
17	88DTG	0	44	-0.32	0.29	0.7	0.7	0.4034
18	10FIRVY	2	217	-0.22	0.21	1.0	1.0	0.4333
19	20IMRTVL	0	115	0.13	0.15	1.0	0.9	0.5338
20	82MLC	10	4	0.47	0.58	0.2	0.2	0.5402
21	63P	0	311	-0.06	0.16	0.6	0.5	0.8915
22	23I	0	4	0.24	0.85	0.3	0.3	0.9133
23	16E	0	9	-0.05	0.50	0.0	0.0	0.9516
24	46ILV	11	143	0.01	0.18	1.0	0.9	0.9516
25	71TVI	2	181	0.02	0.18	1.0	1.0	0.9516
26	84C	10	2	0.15	0.87	0.0	0.0	0.9516

Table 9: Data-adaptively adjusted model-based variable importance estimates ranked by p-value. *The targeted adjustment level is selected based on the asymptotic bias estimate $B_M^{ETA}(\delta)$ for a truncated IPTW estimator. The parameters p and B_{max} are set to 0.05 and 0.25, respectively. The effective adjustment set is selected data-adaptively. δ^t and δ^e give the proportion of potential confounders contained in the targeted and effective adjustment set, respectively. The table also shows the resistance score assigned to a mutation by the Stanford HIVdb scoring system (accessed on 9/1/2007) and the frequency of the mutation among the 401 treatment change episodes.*

Rank	Mutation	Score	Freq	Estimate	SE	δ^t	δ^e	p-value
1	30N	0	45	-0.93	0.23	0.5	0.5	0.0007
2	48VM	10	16	1.00	0.24	0.5	0.5	0.0007
3	50V	20	5	1.67	0.43	1.0	0.9	0.0009
4	54VA	11	84	0.62	0.16	0.8	0.6	0.0009
5	47V	10	17	1.03	0.30	0.5	0.5	0.0025
6	32I	10	21	0.85	0.26	0.5	0.5	0.0043
7	82AFST	20	100	0.46	0.16	0.7	0.6	0.0120
8	54LMST	11	36	0.54	0.19	0.3	0.3	0.0146
9	88S	0	9	-0.74	0.27	0.2	0.0	0.0179
10	73CSTA	2	66	0.52	0.22	0.6	0.6	0.0390
11	88DTG	0	44	-0.57	0.24	0.7	0.7	0.0390
12	33F	5	44	0.53	0.23	0.7	0.7	0.0419
13	24IF	2	16	0.67	0.31	0.1	0.1	0.0642
14	36ILVTA	0	141	0.30	0.15	1.0	0.5	0.0938
15	53LY	3	33	0.41	0.24	0.8	0.7	0.1410
16	84AV	11	73	0.22	0.17	0.5	0.5	0.3073
17	10FIRVY	2	217	-0.16	0.18	1.0	1.0	0.5457
18	20IMRTVL	0	115	0.14	0.15	1.0	0.9	0.5457
19	23I	0	4	0.68	1.02	0.3	0.0	0.6545
20	90M	10	171	0.13	0.19	1.0	1.0	0.6545
21	82MLC	10	4	0.39	0.62	0.2	0.2	0.6567
22	46ILV	11	143	0.06	0.16	1.0	0.8	0.8080
23	71TVI	2	181	0.06	0.15	1.0	1.0	0.8080
24	16E	0	9	-0.05	0.50	0.0	0.0	0.9308
25	63P	0	311	-0.03	0.17	0.6	0.6	0.9308
26	84C	10	2	0.15	1.74	0.0	0.0	0.9308

for. Perhaps it is only when several of these mutations are adjusted for simultaneously that ETA problems arise.

Table 10: Other PI mutations not adjusted for among those mutations statistically significant at the 0.05 level. *Results are based on the non-parametric estimator using a targeted adjustment set selected based on the asymptotic criterion with* $p = 0.05$ *and* $B_{max} = 0.25$. *The effective adjustment set is selected data-adaptively. If the entry in a cell is empty, the variable importance estimate for the mutation in that row was adjusted for the mutation in that column. If the entry is not empty, the mutation in that column could not be adjusted for and the entry shows the sample correlation between the two relevant mutations.*

	30N	32I	47V	50V	54LMST	54VA	82AFST	84AV	88S
30N						-0.12	-0.20	-0.13	
32I			0.62		0.28		0.25		
47V		0.62			0.41		0.08		
48VM	-0.07			0.32	0.16	0.18	0.32		
50V									
54LMST		0.28	0.41			-0.14		0.21	
54VA							0.58		
82AFST	-0.20	0.25				0.58			
84AV	-0.13				0.21				
88S		-0.04	-0.03	-0.02		-0.08	-0.09	-0.03	

7 Discussion

In this paper, we propose a data-adaptive algorithm intended to increase the robustness of variable importance estimation with respect to violations of the ETA assumption. The algorithm is based on one of two identifiability criteria for selecting a targeted adjustment set as well as a mean-squared-error criterion for selecting an effective adjustment set. The data analysis shows very clearly the importance of selecting an appropriate targeted adjustment set as both unadjusted and fully adjusted analyses lead to unsatisfactory results. The fact that the algorithm chose not to adjust for some variables that have quite small marginal correlations with the mutation of interest suggests that serious practical ETA violations may be much more common than previously thought and underscore the need to assess the validity of this assumption. This point is particularly important since many conventional approaches to biomarker discovery such as regression analysis typically do not reveal such problems through sharply inflated standard errors as seen with the non-truncated IPTW and targeted maximum-likelihood estimator, thus not giving the investigator any warning that the parameter of interest may be poorly identified from the observed data.

The data analysis and the simulation study also illustrate the potential gains in efficiency that can be achieved by selecting the effective adjustment set data-adaptively. In the data analysis, the proposed algorithm for selecting both adjustment sets data-adaptively identified a subset of mutations that is in excellent agreement with current understanding of lopinavir resistance, in better agreement, in particular, than previous analyses that categorically excluded other candidate PI mutations from the adjustment set. These findings suggest that variable importance estimation based on data-adaptive selection of the targeted and effective adjustment sets represents a promising new approach for studying the effects of HIV mutations on clinical virologic response to antiretroviral therapy as well as for biomarker discovery in general.

A number of possible extensions of the methodology discussed in this article exist. First, the approach can be applied in a straightforward way to the estimation of causal effects in the point-treatment setting by use of marginal structural models (Robins et al., 2000). In addition, the methodology can be extended to longitudinal data structures. In that case, selection of the targeted adjustment set would need to be based on the parametric-bootstrap approach for estimating the finite-sample bias of the IPTW estimator since closed-form asymptotic bias estimates for truncated IPTW estimators are not currently available for

the longitudinal setting. Selection of the effective adjustment set would still be based on G-computation point estimates that now, however, would also need to be obtained through a Monte-Carlo simulation. Some care would need to be taken in defining a nested sequence of candidate adjustment sets for each time point. It might be preferable to not consider different candidate adjustment sets for different time points, but instead to define identical candidate adjustment sets across time points.

In face of the small marginal correlations between some of the treatment variables and other candidate PI mutations that could not be adjusted for, it might be useful to explore alternative criteria for defining the sequence of candidate adjustment sets. This may be less important in cases in which the ETA assumption is satisfied, but once the adjustment set is large enough to result in an appreciable violation, it might be advantageous to add covariates to the adjustment set directly based on the effect that they would have on the identifiability criterion.

As mentioned in section 4, inference based on the influence curve may be somewhat optimistic in finite samples if the effective adjustment set is selected data-adaptively. Future research is needed to compare inference based on this approach to inference based on an honest bootstrap procedure to quantify the extent to which the use of the former might be problematic. We note, however, that even in situations in which such p-values may be systematically optimistic, they would be still be useful for obtaining a meaningful ranking of the candidate biomarkers.

References

Bembom, O., Petersen, M., Rhee, S.-Y., Fessel, W., Sinisi, S., Shafer, R., and van der Laan M.J. (2007). Biomarker discovery using targeted maximum likelihood estimation: Application to the treatment of antiretroviral resistant hiv infection. Technical Report 221, UC Berkeley Division of Biostatistics Working Paper Series.

Bembom, O. and van der Laan, M. (2007a). Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators. Technical Report XXX, UC Berkeley Division of Biostatistics Working Paper Series.

Bembom, O. and van der Laan, M. (2007b). Estimating the effect of vigorous physical activity on mortality in the elderly based on realistic individualized treatment and intention-to-treat rules. Technical Report 217, UC Berkeley Division of Biostatistics Working Paper Series.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statstical Society: Series B*, 57:289–300.

Liang, K.-Y. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22.

Moore, K. and van der Laan, M. (2007). Covariate Adjustment in Randomized Trials with Binary Outcomes: Targeted Maximum Likelihood Estimation. Technical Report 215, UC Berkeley Division of Biostatistics Working Paper Series.

Neugebauer, R. and van der Laan, M. (2005). Why prefer double robust estimates in causal inference? *Journal of Statistical Planning and Inference*, 129:405–426.

Rhee, S., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D. L., and Shafer, R. W. (2006). Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences*, 103:17355–17360.

Robins, J., Hernán, M., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

Shafer, R., Jung, D., and Betts, B. (2000). Human immunodeficiency virus type 1 reverse transcriptase and protease mutation search engine for queries. *Nature Medicine*, 6(11):1290–1292.

Sinisi, S. and van der Laan, M. (2004). Deletion/Substitution/Addition algorithm in learning with applications in genomics. *Statistical Appliations in Genetics and Molecular Biology*, 3(1):Article 18.

van der Laan, M. (2006). Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1):Article 2.

van der Laan, M. and Robins, J. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer Series in Statistics. Springer Verlag.

van der Laan, M. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):Article 11.

Wang, Y., Petersen, M., Bangsberg, D., and van der Laan, M. (2006). Diagnosing Bias in the Inverse-Probability-of-Treatment-Weighted Estimator Resulting from Violation of Experimental Treatment Assignment. Technical Report 211, UC Berkeley Division of Biostatistics Working Paper Series.

Chapter 8

Case-Control Studies

8.1 *Estimation Based on Case-Control Designs with Known Prevalance Probability*

The following article appears as it was published in the *International Journal of Biostatistics* in 2008, http://www.bepress.com/ijb/vol4/iss1/17/.

It was originally published on the University of California, Berkeley Division of Biostatistics Working Paper Series website in 2008, http://www.bepress.com/ucbbiostat/paper234/.

Estimation Based on Case-Control Designs with Known Prevalence Probability

Mark J. van der Laan
Division of Biostatistics, University of California, Berkeley
laan@berkeley.edu

Abstract

Case-control sampling is an extremely common design used to generate data to estimate effects of exposures or treatments on a binary outcome of interest when the proportion of cases (i.e., binary outcome equal to 1) in the population of interest is low. Case-control sampling represents a biased sample of a target population of interest by sampling a disproportional number of cases. Case-control studies are also commonly employed to estimate the effects of genetic markers or biomarkers on phenotypes.

In this article we present a general method of estimation relying on knowing the incidence probability, conditional on the matching variable if matching is used.

Our general proposed methodology, involving a simple weighting scheme of cases and controls, maps any estimation method for a parameter developed for prospective sampling from the population of interest into an estimation method based on case-control sampling from this population.

We show that this case-control weighting of an efficient estimator for a prospective sample from the target population of interest maps into an efficient estimator for matched and unmatched case-control sampling. In particular, we show how application of this generic methodology provides us with double robust locally efficient targeted maximum likelihood estimators of the causal relative risk and causal odds ratio for regular case control sampling and matched case control sampling.

1 Introduction.

Case-control sampling is an extremely common design used to generate data to estimate effects of exposures or treatments on a binary outcome of interest when the actual population proportion of cases (i.e. binary outcome equal to 1) is small. As a consequence, it is of interest to present estimators of causal effects or variable importance parameters based on case-control data.

1.1 Formulation of case-control estimation problem.

Let's first formulate the statistical problem. For the sake of concreteness and illustration, our formulation will focus on a case-control point treatment data structure with baseline covariates in which one is concerned with estimation of the causal effect or variable importance of the treatment variable on the binary outcome. Our initial formulation will assume that the variables are not subject to missingness or censoring. Our general methods are straightforward extensions and apply to general case control data structures, including censored data structures and time-dependent longitudinal data structures.

Experimental unit of interest. Let $O^* = (W, A, Y) \sim P_0^*$ represent the experimental unit and corresponding distribution P_0^* of interest, consisting of baseline covariates W, a subsequent monitored treatment/exposure variable A, and a "final" binary outcome Y.

Causal or variable importance parameter of interest. Suppose one is concerned with statistical inference regarding a particular euclidean valued variable importance or causal effect parameter $\psi_0^* = \Psi^*(P_0^*) \in \mathbb{R}^d$ of this distribution P_0^*. For example, one might be interested in the marginal causal additive effect of a binary treatment $A \in \{0, 1\}$ defined as

$$\begin{aligned} \psi_0^* &\equiv E_0^*\{E_0^*(Y \mid A = 1, W) - E_0^*(Y \mid A = 0, W)\} = E_0^*(Y_1) - E_0^*(Y_0) \\ &= P_0^*(Y_1 = 1) - P_0^*(Y_0 = 1), \end{aligned}$$

where the latter causal effect interpretation of this parameter of P_0^* requires the notion of treatment specific counterfactual outcomes Y_0, Y_1, viewing $(W, A, Y = Y_A)$ as a time-ordered missing data structure on the full data structure (W, Y_0, Y_1), and one needs to assume the randomization assumption stating that A is independent of Y_0, Y_1, given W. The latter causal parameter formulation ψ_0^* can also be viewed as a W-adjusted variable importance (of variable A) parameter of the true regression of Y on A, W, in which case there is no need

to assume the time ordering ($W \Rightarrow A \Rightarrow Y$), the missing data structure assumption, or the randomization assumption, and the adjustment set W is user supplied (and does thus not need to correspond with the set of all confounders of A): see van der Laan (2006) for a general formulation of variable importance parameters and its direct relation to causal effect parameters.
One can also define the parameter of interest as a causal relative risk

$$\psi_0^* = \frac{E_0^* E_0^*(Y \mid A=1, W)}{E_0^* E_0^*(Y \mid A=0, W)} = \frac{EY_1}{EY_0} = \frac{P(Y_1=1)}{P(Y_0=1)},$$

or a causal odds ratio,

$$\psi_0^* = \frac{P(Y_1=1)P(Y_0=0)}{P(Y_1=0)P(Y_0=1)},$$

or their variable importance analogue.

We will use these particular marginal causal effects or marginal variable importance parameters as our main examples in order to illustrate our proposed methodology for case-control data, including our proposed targeted maximum likelihood estimation methodology.

Model for target probability distribution. A model for O^* is obtained by modelling this distribution of O^*: for example, one might know that A is independent of W, one might know the actual distribution (treatment mechanism) $P_0^*(A=a \mid W)$, or one might assume a marginal structural model

$$E_0^*(Y_a \mid V) = E_0^*(E_0^*(Y \mid A=a, W) \mid V) = m(a, V \mid \beta_0^*),$$

where $V \subset W$ denotes some user supplied potential effect modifier of interest, and $m(\cdot \mid \beta)$ some parameterization modelling the causal effect of the intervention $A=a$ on the outcome Y, conditional on V. If one wishes to avoid making causal assumptions, the marginal structural parameter represents the effect of a change in variable A on the mean outcome of Y within subgroups $V=v$, controlling for potential confounders W. We will denote such a model for P_0^* with $\mathcal{M}^*$: i.e., it is assumed that $P_0^* \in \mathcal{M}^*$.

Case-control sampling and its probability distribution. If one would sample n i.i.d. observations $O_1^*, \ldots, O_n^* \sim P_0^*$, then we could (e.g.) apply the locally efficient targeted MLE of ψ_0^* (see e.g. van der Laan and Rubin (2006) or Moore and van der Laan (2007)), or one could use double robust estimating function methodology (van der Laan and Robins (2002)).

However, this so called prospective sampling scheme is often considered impractical and ineffective in situations in which the probability $P_0^*(Y = 1)$ on the event $Y = 1$ (say disease) is very small. For example, if the proportion of diseased in the population of interest is one in hundred thousand, then one would have to sample millions of observations in order to have some cases (i.e, $Y_i = 1$) in the sample. This sparsity of cases in the population of interest is precisely the typical motivation for case-control sampling.

We will distinguish between two types of case-control sampling: independent or un-matched case-control sampling and matched case-control sampling. In both cases, the marginal distribution of the cases and the marginal distribution of the controls is completely determined by the population (i.e. prospective sampling) distribution P_0^* of the random variable (W, A, Y) of interest.

Independent Case-Control Sampling. One first samples a *case* by sampling (W_1, A_1) from the conditional distribution of (W, A), given $Y = 1$. Subsequently, one samples J *controls* (W_0^j, A_0^j) from the conditional distribution of (W, A), given $Y = 0$, $j = 1, \ldots, J$. It is allowed that these J control observations are dependent as long as their marginal distributions are indeed equal to the conditional distribution of W, A, given $Y = 0$.

This results in an experimental unit observed data structure:

$$O = ((W_1, A_1), (W_0^j, A_0^j : j = 1, \ldots, J)) \sim P_0,$$

where we denote the sampling distribution of this data structure O described above with P_0. Thus, a case control data set will consists of n independent and identically distributed observations $O_1, \ldots, O_n$ with sampling distribution P_0 described above. That is, we treat the cluster consisting of one case and J controls as the experimental unit, and the marginal distribution of the case and controls are specified as above by P_0^*.

Matched Case-Control Sampling. One specifies a categorical matching variable $M \subset W$. One first samples a case by sampling (M_1, W_1, A_1) from the conditional distribution of (M, W, A), given $Y = 1$. Subsequently, one samples J controls (M_0^j, W_0^j, A_0^j) from the conditional distribution of (M, W, A), given $Y = 0, M = M_1$. That is, with probability equal to 1 we have $M_0^j = M_1$, $j = 1, \ldots, J$. It is allowed that these J control observations are dependent as long as their marginal distributions are indeed equal to the conditional distribution of M, W, A, given $Y = 0, M = M_1$.

This results in an experimental unit data structure:

$$O = ((M_1, W_1, A_1), (M_0^j = M_1, W_0^j, A_0^j : j = 1, \ldots, J)) \sim P_0,$$

where we denote the sampling distribution of this data structure O described above with P_0. Thus, a matched case-control data set will consists of n independent and identically distributed observations $O_1, \ldots, O_n$ with sampling distribution P_0 described above. That is, we treat the cluster consisting of one case and the J matched controls as the experimental unit, and the marginal distribution of the case and J controls are specified as above by P_0^*

We will also refer to the independent case-control experiment and the matched case-control experiments as Case-Control Design I and Case-Control Design II, respectively.

Extensions. Our methods naturally handle the case that J is random and thus varies per experimental unit, assuming that the marginal distributions of cases and controls, conditional on $J = j$, do not depend on j. In the situation that a case was never coupled to a set of controls one can artificially create such couplings, and apply our methods, and one could average over a variety of sensible coupling schemes. The latter shows that if the true independent case control design simply involves sampling a set of cases and an independent set of controls, without any coupling, then our case control weighting methods show that one should weight each case by q_0 and each control by $(1 - q_0)/\bar{J}$, where $\bar{J}$ is the number of controls divided by the number of cases. In the discussion we show the simple extension of our methods to some variations on these case-control designs I and II, such as pair-matched case-control designs, case-control sampling within strata, and counter-match case control designs. We also note here that our sampling model for O^* corresponds with sampling with replacement from a particular population with population distribution P_0^*. Such a model is appropriate if the size of the total population is large relative to sample size n.

The estimation problem. The statistical problem is now to estimate the parameter $\psi_0 = \Psi^*(P_0^*)$ of the population distribution $P_0^* \in \mathcal{M}^*$ of (W, A, Y), known to be an element of some specified model $\mathcal{M}^*$, based on the case-control data set $O_1, \ldots, O_n \sim P_0$.

Known or sensitivity analysis parameters/weights. We define

$$q_0 \equiv P_0^*(Y = 1) \text{ and } q_0(\delta \mid M) \equiv P_0^*(Y = \delta \mid M),$$

as the marginal probability of being a case, and the conditional probability of being a case/non-case, conditional on the matching variable. It is assumed that

these probabilities are between 0 and 1. In addition, we define the quantity

$$\bar{q}_0(M) \equiv q_0 \frac{P_0^*(Y=0 \mid M)}{P_0^*(Y=1 \mid M)} = q_0 \frac{q_0(0 \mid M)}{q_0(1 \mid M)}.$$

We note that $\bar{q}_0(M)$ is determined by q_0 and $q_0(1 \mid M) = P_0^*(Y=1 \mid M)$, and we also note that $E_0 \bar{q}_0(M_1) = 1 - q_0$. These two quantities q_0 and $\bar{q}_0(M)$ (for matched case-control studies) will be used to weight the cases and controls to obtain valid estimation procedures.

In order to be able to identify the wished causal parameters, for case-control design I, we only need to assume q_0 is known, and, for matched case-control design II, we assume q_0 and $\bar{q}_0(m)$ for each m are known. However, we note here that for matched case-control designs one can also assume that q_0 and

$$r_0(m) \equiv P_0^*(Y=0, M=m)$$

(instead of $\bar{q}_0(1 \mid m)$) are known We note that, given $r_0(m)$, $\bar{q}_0(m)$ is known up till a simple to estimate nuisance parameter $P(M_1 = m)$:

$$\bar{q}_0(m) = \frac{r_0(m)}{P_0(M_1 = m)}.$$

As a consequence, our case-control weighted estimation procedures using q_0, $\bar{q}_0(m)$ still apply in settings in which one assumes q_0 and $r_0(m)$ are known, by replacing $\bar{q}_0(m)$ by its estimate $\frac{r_0(m)}{\frac{1}{n}\sum_{i=1}^n I(M_{1i}=m)}$.

Observed data model. In this article, we will assume that q_0 is known, and that, for matched case-control designs we also assume that $\bar{q}_0(M)$, or equivalently, $q_0(1 \mid m) = P_0^*(Y=1 \mid M=m)$ is known for each m. In our accompanying technical report we show that if the "treatment mechanism" $g_0^*(a \mid w) = P_0^*(A=a \mid W=w)$ is known, as it would be in a case control study nested in a randomized trial, then we can estimate the relative risk or odds ratio parameters without a need to know (any of) q_0 or $\bar{q}_0(M)$.

The model $\mathcal{M}^*$, possibly including the knowledge q_0 or $\bar{q}_0(M)$, imply now models for the marginal distribution of the cases (M_1, W_1, A_1) and the marginal distributions of the controls (M_1, W_2^j, A_2^j), $j = 1, \ldots, J$. The model $\mathcal{M}^*$ does not imply much, if anything, about the dependence structure among (M_1, W_1, A_1), (M_1, W_2^j, A_2^j), $j = 1, \ldots, J$, beyond the fact that, for matched case-control studies, all its components (i.e., the case and control observations) share a common variable M_1. Let $\mathcal{M}$ be the model for the observed

data distribution P_0 compatible with $\mathcal{M}^*$ (i.e., its marginals are specified by P_0^*).

One possible and probably very common model $\mathcal{M}$ is to assume that, given the first draw (M_1, W_1, A_1) from (M, W, A), given $Y = 1$, the control observations are all *independent* draws from the specified conditional distributions. Note that in this latter model the marginal distributions for the case and control observations implied by P^* describe now the whole case-control sampling distribution P, so that we can write $\mathcal{M} = \{P(P^*) : P^* \in \mathcal{M}\}$, where $P(P^*)$ is the distribution of O implied by P^*.

Other possible models might specify in another manner, or not specify at all, the dependence structure and could, for example, be represented as $\{P(P^*, \eta) : P^* \in \mathcal{M}^*, \eta\}$, where the nuisance parameter η in combination with P^* describes the complete joint distribution of case and control observations $(M_1, Z_1), (M_1, Z_2^j : j = 1, \ldots, J)$ compatible with its marginal distributions implied by P^*.

We note that knowing q_0 does not put restrictions on the data generating distribution P_0 since one conditions on $Y = 1$, but for case-control design I it does allow identification of the wished parameters by expressing them as a function of the distribution of the observed case-control data-structure and q_0. Similarly, for matched case-control designs, knowing q_0 and $r_0(\cdot)$ does not put restrictions on the data generating distribution P_0 for matched case-control designs, but it allows one to express the wished parameter as a function of the distribution of the data and (q_0, r_0). It remains to be investigated if knowing q_0 and $\bar{q}_0$ puts a restriction on the data generating distribution for matched-case-control designs.

1.2 Overview of article.

In Section 2 we present our general solution to the estimation problem for these two types of case control designs I and II, which weights the cases and controls with q_0 and $(1 - q_0)/J$ ($\bar{q}_0(M)/J$ for case control design II), respectively, and then applies a method developed for prospective sampling to estimate the parameter of interest (e.g., targeted maximum likelihood estimators or estimating equations for the causal effect or variable importance parameter ψ_0 of interest), as if the data was directly drawn from the population distribution P_0^* of interest. In other words, each estimating function for ψ_0^* or likelihood for P_0^* in the underlying model $\mathcal{M}^*$ maps into a "case-control"-weighted estimating function or likelihood for the observed data model $\mathcal{M}$ (whatever nuisance parameter specification $P(P^*, \eta)$ it might have beyond the description of its marginal distributions in terms of P^*).

Beyond the weighting, we point out that one should aim to select the best among these case-control weighted estimating equations/procedures for the observed case-control data. In Section 3 we show the important and convenient result that case-control weighting of the efficient procedure for the parameter of interest (as formalized by the efficient influence curve) in the prospective sampling model $\mathcal{M}^*$ maps into the efficient procedure for the observed case-control data model $\mathcal{M}$. This implies, in particular, that case-control weighting of the locally efficient targeted maximum likelihood estimator developed for prospective sampling model $\mathcal{M}^*$ results in a locally efficient targeted maximum likelihood estimation procedure for case-control sampling. In general, the power of our generic method is that one can map the estimation procedures developed for prospective sampling into highly or fully efficient estimation procedures for case-control sampling. In particular, our method is now able to fully exploit software developed for prospective sampling.

To summarize, in Section 2 and Section 3 we establish general properties of our case-control weighted mapping from estimating functions/influence curves/gradients for the parameter of interest for model $\mathcal{M}^*$ into estimating functions/influence curves/gradients for the parameter of interest for the observed data model $\mathcal{M}$, showing that 1) the case-control weighting does map each parameter-specific influence curve for the model $\mathcal{M}^*$ into a parameter-specific influence curve for model $\mathcal{M}$, 2) it maps the efficient influence curve/ canonical gradient for model $\mathcal{M}^*$ into the efficient influence curve/canonical gradient for model $\mathcal{M}$, and 3) that our case-control weighting inherits any robustness of estimating functions/influence curves for model $\mathcal{M}^*$.

We suggest that even in cases that q_0 (or $q_0(1 \mid M)$ for matched case control designs) is unknown, it is of interest to present these estimators and inferences for an interval of possible q_0-values, thereby presenting a sensitivity analysis.

As an example we show that indeed for case-control design I the case-control weighted targeted maximum likelihood estimator is indeed a locally efficient double robust estimator. This implementation of a targeted maximum likelihood estimators needs to guarantee that the initial maximum likelihood fit of the logistic regression $P_0^*(Y = 1 \mid A, W)$ is proportional to q_0, which is a requirement for these double robust estimators to *not* suffer from a large variance due to the singularity $q_0 \approx 0$. The latter is precisely guaranteed by our case-control weighting method.

These double robust targeted maximum likelihood estimators rely on knowing the incidence probability q_0 and, for case-control design II, $\bar{q}_0(M)$, beyond either a correctly specified model for $Q^*(A, W) = P_0^*(Y = 1 \mid A, W)$ or a correctly specified model for $g_0^*(a \mid W) = P_0^*(A = a \mid W)$.

In Section 4, we end this article with a discussion and point out a number

of extensions. Various technical proofs are deferred to the Appendix.

1.3 Some relevant literature.

Case-control studies are probably one of the most commonly used designs, if not the most used design. For example, searching for case-control analysis on PubMed resulted in a list of 56,000 articles. Their use is not limited to public health applications; case-control studies are also frequently performed in econometric applications (See Manski and Lerman (1977), Manski and McFadden (1981), Cosslett (1981)). Logistic regression is the most commonly used model in the literature for case-control studies. Conditional logistic regression is the prominent method in the literature for matched case-control studies and the statistical methodology goes back to the early 80's.

We will discuss these two methods briefly as well as related IPTW methods, as it goes without saying that an overview of the literature in this area is not possible. However, our proposed general methodology is not covered by the current literature, as far as we know.

Some of the key papers on logistic regression in standard case-control studies are Anderson (1972), Prentice and Pyke (1979), Breslow (1996), and Breslow and Day (1980). Breslow et al. (2000) establish asymptotic efficiency of the standard maximum likelihood estimator ignoring the case-control sampling. The most frequently cited sources for conditional logistic regression for matched case-control studies are Breslow and Day (1980), Holford et al. (1978), and Breslow et al. (1978). Various books considering case-control studies are Schlesselman (1982), Collett (1991), Jewell (2004), Rothman and Greenland (1998), and Hosmer and Lemeshow (2000), among others.

Cohort studies differ from case-control studies in that they sample exposed ($A = 1$) and unexposed ($A = 0$) individuals rather than diseased ($Y = 1$) and non-diseased ($Y = 0$). When cohort studies are matched, they are matched based on the exposure variable in an effort to reduce the bias found in observational studies. There has been much work in this area, particularly in the analysis and matching of cohort studies, by W.G. Cochran, D.B. Rubin, P.R. Rosenbaum, and N. Thomas. A collection of this work can be found in Rubin (2006). A thorough discussion of cohort study design can also be found in Rothman and Greenland (1998).

The method of adding an intercept to a standard logistic regression fit, and, in that manner, estimating effects different from the odds-ratio has been presented in the literature (see e.g. Anderson (1972), Prentice and Breslow (1978), Greenland (1981), Morise et al. (1996), Wachholder (1996), Greenland (2004)).

Matched case-control studies are most frequently handled with conditional logistic regression models, but these designs and methods also have limitations. Firstly, it does not allow estimation of the effect of the matching variable on the diease (see, Schlesselman (1982), Rothman and Greenland (1998)): Any variable used for matching cannot be studied as a risk factor, since cases and controls are constrained to be equal with respect to the variables that are matched. Secondly, matching can hurt the precision if the matching variable is correlated with the exposure variable and not disease, which is often called over-matching. Finally, as we remarked from the start, these methods are by necessity heavily model based, while the methods presented here, relying on knowing the case-control weights, allow double robust locally efficient estimation in semiparametric models, thereby allowing the use of methods which minimize the reliance of the inference on unknown model assumptions.

Robins (1999) discusses the approximately correct IPTW-method for estimation of the unknown parameters in a marginal structural logistic regression model for a direct effect analysis based on standard case-control data under the assumption that the population proportion of cases, q_0, is small. We also refer to Newman (2006) for an IPTW-type approach for fitting marginal structural models based on case-control data. Mansson et al. (2007) investigate a variety of IPTW and propensity score methods in case-control studies through a simulation study, which includes the IPTW estimator for the logistic marginal structural model.

Notation. We introduce now some useful notation. Let $O^* \rightarrow D^*(O^*)$ represent an estimating function or loss function for O^* that can thus be used to estimate the parameter of interest of P_0^* based on an i.i.d sample of O^*. This article is concerned with mapping this function D^* into an estimating function of loss function for this same parameter of interest, but now based on sampling O (i.e., a biased sample for O^*). Given such a function $D^*(O^*)$, we define a case-control weighted version $D_{q_0}(O) \equiv q_0 D^*(W_1, A_1, 1) + \frac{1}{J}\sum_{j=1}^{J} \bar{q}_0(M_1) D^*(W_2^j, A_2^j, 0)$ of D^*, which is now a function of the observed experimental unit O. We define the expectation operator $P_{0,q_0} D^* = P_0 D_{q_0}$, which thus simply takes the expectation of the case-control weighted function $D_{q_0}(O)$ w.r.t. P_0. Similarly, we define the empirical expectation $P_{n,q_0} D^* = P_n D_{q_0}$ as the empirical mean of the case-control weighted D_{q_0}, where P_n is the empirical distribution of $O_1, \ldots, O_n$. We apply this notation to both case-control designs, where for case-control design I $\bar{q}_0(M_1)$ reduces to $1 - q_0$.

2 Case-Control weighting of estimation procedures developed for prospective sampling.

Throughout this section, we will make the convention that $\bar{q}_0(M)$ reduces to $1 - q_0$ in the case control design I, so that we can state our results for both the regular case-control design I and the matched case-control design II in one formula.

We start out with stating the theorem which proves that the case-control weighting maps a function of O^* into a function of the case-control data structure O, while preserving the expectation of the function.

Definition 1 *(**Case-control weighted function**) Given a $D^*(O^*) = D^*(W, A, Y)$ we define the case-control weighted version of D^* as*

$$D_{q_0}(O) \equiv q_0 D^*(M_1, W_1, A_1, 1) + \frac{1}{J}\sum_{j=1}^{J} \bar{q}_0(M_1) D^*(M_1, W_2^j, A_2^j, 0),$$

where in the special case of Case Control Design I, we have $\bar{q}_0(M) = 1 - q_0$.

Theorem 1 *(**Unbiased estimating function mapping**) Let $D^*(O^*) = D^*(W, A, Y)$ be a function so that $P_0^* D^* \equiv E_{P_0^*} D^*(O^*) = 0$. Then $P_0 D_{q_0} = 0$. In particular, in Case Control Design I,*

$$D_{q_0}(0) \equiv q_0 D^*(W_1, A_1, 1) + (1 - q_0)\frac{1}{J}\sum_{j=1}^{J} D^*(W_2^j, A_2^j, 0)$$

satisfies $P_0 D_{q_0} = 0$.

In more generality, for any function D^* and corresponding case control weighted function D_{q_0}, we have

$$P_0 D_{q_0} = P_0^* D^*.$$

Proof. We provide the proof for case-control design II and we suppress the index q_0 in D_{q_0}. The same proof applies to case-control design I. First, we note that $P_0 q_0 D(M_1, W_1, A_1, 1) = \int_{M_1, W_1, A_1} D(M_1, W_1, A_1, 1) P_0^*(M_1, W_1, A_1, Y = 1)$. Secondly, we note that

$$P_0 \bar{q}_0(M_1) D(M_1, W_2^j, A_2^j, 0) =$$
$$\int_{m,w,a} D(m, w, a, 0) \bar{q}_0(m) P_0(M_1 = m) P_0^*(W = w, A = a \mid M = m, Y = 0),$$

where we also need to note that $P_0(M_1 = m) = P_0^*(M = m \mid Y = 1)$. We have

$$\begin{aligned}
&\bar{q}_0(m)P_0(M_1 = m)P_0^*(W = w, A = a \mid M = m, Y = 0) \\
&= \frac{\bar{q}_0(m)P_0^*(M=m|Y=1)P_0^*(W=w,A=a,M=m,Y=0)}{P_0^*(Y=0,M=m)} \\
&= P_0^*(M = m, W = w, A = a, Y = 0).
\end{aligned}$$

This proves that

$$\begin{aligned}
&P_0 D = \textstyle\int_{M_1,W_1,A_1} D(M_1, W_1, A_1, 1)P_0^*(M_1, W_1, A_1, Y = 1) \\
&+\frac{1}{J}\textstyle\sum_{j=1}^{J} \int_{M_1,W_2,A_2} D(M_1, W_2, A_2, 0)P_0^*(M_1, W_2, A_2, Y = 0) \\
&= P_0^* D = 0.
\end{aligned}$$

This completes the proof. □

In the next section we establish general properties of this mapping which help us to understand the generality and optimality of the statistical approach for dealing with case-control sampling implied by this mapping. In this section we focus on the statistical (i.e., methodological) implications of this mapping for the analysis of case-control data,

2.1 Preservation of robustness of case-control weighted functions.

If a function D^* satisfying $P_0^* D(P_0^*) = 0$ also satisfies the robustness property $P_0^*(D(P^*)) = 0$ for any $P^* \in \mathcal{M}_1^* \subset \mathcal{M}^*$ for a submodel $\mathcal{M}_1^*$, then the same robustness w.r.t. to misspecification of P_0^* applies to D_{q_0} since, for $P^* \in \mathcal{M}_1^*$, $P_0 D_{q_0}(P^*) = P_0^* D(P^*) = 0$.

In particular, double robust estimating functions for censored and causal inference data structures and models $\mathcal{M}^*$, as presented in general in van der Laan and Robins (2002), are mapped into double robust case-control weighted estimating functions.

In the remainder of this section we outline the general statistical methods implied by the case-control weighted mapping. Estimating function methodology developed for prospective sampling immediately implies now, through the case-control weighted mapping, estimating function methodology for case-control sampling. In particular, in view of the general estimating function theory presented in van der Laan and Robins (2002) it follows that the case control mapping is a mapping from estimating functions (or gradients, see van der Laan and Robins (2002)) developed for a model for P_0^* into estimating functions based on case-control sampling from P_0. For details we refer to our technical report, and here we suffice with an illustration.

2.2 Example: Case-control weighted double robust estimating function.

Let's illustrate this estimating function method by constructing a double robust estimator of the additive causal effect $\psi_0^* = E(Y_1 - Y_0)$ for a nonparametric model $\mathcal{M}^*$ for the distribution P_0^* of (W, A, Y). Let $g_0^*(A \mid M, W)$ denote the conditional distribution of A, given W, and let $Q_0^*(M, W, A)$ denote the conditional probability of Y, given M, W, A, under P_0^*.

The double robust efficient estimating function for sampling from P_0^* is given by

$$\begin{aligned} D^*(\psi^*, g^*, Q^*)(O^*) &= \left\{ \frac{I(A=1)}{g^*(1 \mid M, W)} - \frac{I(A=0)}{g^*(0 \mid M, W)} \right\} (Y - Q^*(M, W, A)) \\ &\quad + Q^*(M, W, 1) - Q^*(M, W, 0) - \psi^*, \end{aligned} \tag{1}$$

where g^* and Q^* represent candidates for the nuisance parameters g_0^* and Q_0^* of this estimating function for ψ_0^*.

It is double robust in the sense that

$$E_0^* D^*(\psi_0^*, g^*, Q^*)(O^*) = 0 \text{ if either } g^* = g_0^* \text{ or } Q^* = Q_0^*,$$

and in both cases one needs that $g^*(1 \mid W) g^*(0 \mid W) > 0$ a.e. Let $D^*(g^*, Q^*)$ be defined so that $D^*(\psi^*, g^*, Q^*) = D^*(g^*, Q^*) - \psi^*$.

The weighted double robust estimating function for case-control data is thus given by:

$$\begin{aligned} D_{q_0}(\psi^*, g^*, Q^*)(O) &= q_0 D^*(\psi^*, g^*, Q^*)(M_1, W_1, A_1, 1) \\ &\quad + \frac{\bar{q}_0(M_1)}{J} \sum_{j=1}^{J} D^*(\psi^*, g^*, Q^*)(M_1, W_2^j, A_2^j, 0), \end{aligned}$$

or we can define it as

$$\begin{aligned} D_{q_0}(\psi^*, g^*, Q^*)(O) &= q_0 D^*(g^*, Q^*)(M_1, W_1, A_1, 1) \\ &\quad + \frac{\bar{q}_0(M_1)}{J} \sum_{j=1}^{J} D^*(g^*, Q^*)(M_1, W_2^j, A_2^j, 0) - \psi^*. \end{aligned}$$

This estimating function is now also double robust for case control data:

$$E_0 D_{q_0}(\psi_0^*, g^*, Q^*) = 0 \text{ if either } g^* = g_0^* \text{ or } Q^* = Q_0^*,$$

and in both cases one needs that $g^*(1 \mid W) g^*(0 \mid W) > 0$ a.e.

The solution ψ_n of the case-control weighted estimating equation:

$$P_n D_{q_0}(g_n^*, Q_n^*) - \psi^* = 0$$

exists in closed form and is given by:

$$\begin{aligned}\psi_n &= \frac{1}{n}\sum_{i=1}^{n} q_0 D^*(g_n^*, Q_n^*)(M_{1i}, W_{1i}, A_{1i}, 1) \\ &\quad + \frac{\bar{q}_0(M_{1i})}{J}\sum_{j=1}^{J} D^*(g_n^*, Q_n^*)(M_{1i}, W_{2i}^j, A_{2i}^j, 0).\end{aligned}$$

This estimator is now consistent if either g_n^* consistently estimates g_0^* or Q_n^* consistently estimates Q_0^*, which explains why it is called double robust.

Under some extra appropriate regularity conditions, this estimator is also asymptotically linear and thereby has a normal limit distribution (see van der Laan and Robins (2002) for general "central limit" theorems for solutions of estimating equations). In particular, if g_n^* consistently estimates g_0^* and Q_n^* consistently estimates Q_0^*, then, under appropriate regularity conditions, ψ_n is asymptotically linear with influence curve $D_{q_0}(g_0^*, Q_0^*, \psi_0)$ and is thus asymptotically efficient. The estimators g_n^* and Q_n^* can be based on case-control weighting of maximum likelihood estimators for the prospective model, as presented in next subsection.

Statistical behavior of double robust estimator when cases are rare. Inspection of this influence curve D_{q_0} sheds some light on the statistical behavior of this double robust estimator for the important case that $q_0 \approx 0$ is very small. In particular, we are interested in how well one can estimate the relative effect ψ_0/q_0, since ψ_0 is itself very small. It follows that, in general, the influence curve of ψ_n/q_0 as an estimator of ψ_0/q_0 will blow up for small values q_0, *except if it guaranteed that* $Q_n^* = q_0 Q_n^{\#}$ *for some bounded estimator* $Q_n^{\#}$. Therefore, in our proposed targeted maximum likelihood or double robust estimator we propose such estimators based on either case-control weighted logistic regression fits or intercept adjusted logistic regression fits (see Section 2 accompanying technical report).

2.3 Case-control weighted loss functions.

Our case-control weighting can also be used to map loss functions for the underlying model $\mathcal{M}^*$ into loss functions for the observed data model $\mathcal{M}$. In particular, we can construct a case-control weighted log likelihood loss function.

Theorem 2 *(**Case Control Weighted Log-Likelihood Loss function**) Define the following case-control weighted log-likelihood loss function for the density p_0^* of O^* under sampling of $O \sim P_0$:*

$$L(p^*, O) = q_0 \log p^*(M_1, Z_1, 1) + \bar{q}_0(M_1)\frac{1}{J}\sum_{j=1}^{J} \log p^*(M_1, Z_2^j, 0).$$

In particular, in Case Control Design I, we have

$$L(p^*, O) = q_0 \log p^*(M_1, Z_1, 1) + (1 - q_0)\frac{1}{J}\sum_{j=1}^{J} \log p^*(M_1, Z_2^j, 0).$$

We have

$$p_0^* = \arg\max_{p^*} E_0 L(p^*, O),$$

where the argmax is taken over all densities p^*. That is, the density maximizing the expectation of the loss function $L(p^*, O)$ is unique and given by the density p_0^* of O^*.

The proof of this theorem is similar to the proof of Theorem 1 and is therefore omitted.

2.4 Case-control weighted maximum likelihood estimation.

Given a specified model $\mathcal{M}^*$ for p_0^*, we can estimate P_0^* with the case-control weighted maximum likelihood estimator:

$$p_n^* = \arg\max_{p^* \in \mathcal{M}^*} \sum_{i=1}^{n} L(O_i, p^*).$$

The implementation of this weighted maximum likelihood estimator simply involves assigning weights q_0 to the cases, assigning weights $\bar{q}_0(M_{1i})/J$ to the corresponding J controls, and then implementing the maximum likelihood estimator for prospective sampling (i.e. treating the sample of cases and controls as an i.i.d sample of P_0^*), thus ignoring the case control sampling.

For example, let's consider the point treatment data structure $O^* = (M, W, A, Y)$. Consider a nonparametric model for the marginal distribution of W, Q_W^*, a model $\{g_\eta^* : \eta\}$ for $g_0^*(A \mid M, W)$, and a model $\{Q_\theta^* : \theta\}$ for the conditional distribution $P_0^*(Y = 1 \mid M, W, A) = Q_0^*(M, W, A)$.

The case-control weighted maximum likelihood estimator of the marginal distribution of W is now the weighted empirical distribution of the pooled sample $(W_{1i}, (W_{2i}^j : j = 1, \ldots, J))$. Similarly, the case-control weighted maximum likelihood estimator of $g_0^*(A \mid W)$ is given by

$$\eta_n = \arg\max_{\eta} \sum_{i=1}^{n} q_0 \log g_\eta^*(A_{1i} \mid M_{1i}, W_{1i}) + \frac{\bar{q}_0(M_{1i})}{J} \sum_{j=1}^{J} \log g_\eta^*(A_{2i}^j \mid M_{1i}, W_{2i}^j),$$

and the case-control weighted maximum likelihood estimator of $Q_0^*(M, W, A)$ is given by

$$\theta_n = \arg\max_{\theta} \sum_{i=1}^{n} q_0 \log Q(M_{1i}, W_{1i}, A_{1i}) + \frac{\bar{q}_0(M_{1i})}{J} \sum_{j=1}^{J} \log(1 - Q(M_{1i}, W_{2i}^j, A_{2i}^j)).$$

Indeed, it follows that each of these case-control weighted maximum likelihood estimators can be implemented by assigning the two weights q_0 and $\bar{q}_0(M_1)$ to the cases and controls, respectively, and apply the standard maximum likelihood estimator of the density p_0^* under prospective sampling.

Given the weighted maximum likelihood estimators Q_{1n}^* and Q_n^*, described above, the corresponding substitution estimator of $EY_a = E_{Q_1^*} Q^*(W, a)$ is given by

$$\psi_n(a) = \frac{1}{\sum_{i=1}^{n} \{q_0 + \bar{q}_0(M_{1i}))\}} \sum_{i=1}^{n} q_0 Q_n^*(M_{1i}, W_{1i}, a) + \frac{\bar{q}_0(M_{1i})}{J} \sum_{j=1}^{J} Q_n^*(M_{1i}, W_{2i}^j, a).$$

In particular, these estimators of EY_0 and EY_1 now map into an estimator $\psi_n(1)/\psi_n(0)$ of the relative risk EY_1/EY_0.

2.5 Case-control weighted targeted maximum likelihood estimation.

Targeted maximum likelihood estimation is a general methodology introduced in van der Laan and Rubin (2006) and illustrated with a variety of examples. The case-control weighting allows us now to provide a case-control weighted targeted maximum likelihood estimation methodology targeting the parameter of interest.

Specifically, let $D^*(P_0^*)$ be the efficient influence curve of the parameter $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$. Consider an initial estimator P_n^{*0} of P_0^* based on $O_1, \ldots, O_n$ such as a case-control weighted maximum likelihood estimator according to a working model within $\mathcal{M}^*$. Let $\{P_n^*(\epsilon) : \epsilon\}$ be a submodel of $\mathcal{M}^*$ with

parameter ϵ satisfying that the linear span of its score at $\epsilon = 0$ includes $D^*(P_n^{*0})$. Let ϵ_n^1 be the case-control weighted maximum likelihood estimator of ϵ:

$$\epsilon_n^1 = \arg\max P_{n,q_0} \log p_n^{*0}(\epsilon).$$

This yields an update $P_n^{*1} = P_n^{*0}(\epsilon_n^1)$ of the initial estimator P_n^{*0}. We iterate this updating process till step k at which $\epsilon_n^k \approx 0$ and we denote the final update with P_n^*. By the score condition, this final estimator solves the case-control weighted efficient influence curve:

$$0 = P_{n,q_0} D^*(P_n^*) = P_n D_{q_0}(P_n^*)$$

up till numerical precision (see van der Laan and Rubin (2006)). We refer to $\psi_n = \Psi^*(P_n^*)$ as the case-control weighted targeted maximum likelihood estimator of ψ_0.

One particular approach for establishing the asymptotics of this estimator is obtained under the assumption that $D^*(P^*) = D^*(\psi^*, \eta^*)$ for some nuisance parameter, thereby assuming an estimating function representation for the efficient influence curve. (This assumption is not necessary at all to establish the same asymptotics: see van der Laan and Rubin (2006).) In this case, it follows that the targeted maximum likelihood estimator ψ_n solves $P_n D_{q_0}(\psi_n, \eta_n^*) = 0$ so that one can establish asymptotic linearity of ψ_n and derive its influence curve under relatively standard differentiability and empirical process conditions.

In particular, if η_n^* is a consistent estimator of a η_0^* satisfying $P_0 D_{q_0}(\psi_0, \eta_0^*) = 0$, then under such standard conditions, asymptotic consistency and asymptotic linearity can be established. For example, if $\eta_0^* = \eta(P_0^*)$ is the true parameter, then ψ_n will have influence curve given by $D_{q_0}(\psi_0, \eta_0^*)$.

2.6 Case-control weighted targeted MLE of marginal causal effect for case control data.

We will illustrate the targeted maximum likelihood estimator for the parameter $\psi_0 = EY_1 - EY_0$ and the nonparametric model $\mathcal{M}^*$ for the point treatment data structure $(W, A, Y) \sim P_0^*$.

Recall that the double robust estimating function/efficient influence curve of Ψ under i.i.d sampling from P_0^* is given by

$$\begin{aligned} D^*(g^*, Q^*)(M, W, A, Y) &= \left\{ \frac{I(A=1)}{g^*(1 \mid M, W)} \frac{I(A=0)}{g^*(0 \mid M, W)} \right\} \\ &\quad \times (Y - Q_2^*(M, W, A)) \end{aligned}$$

$$\begin{aligned} & +Q_2^*(M,W,1) - Q_2^*(M,W,0) - \Psi(Q^*) \\ \equiv \; & D_1^*(g^*,Q^*)(M,W,A,Y) + D_2^*(Q^*)(M,W), \end{aligned}$$

where $Q^* = (Q_1^*, Q_2^*)$ represents both the marginal distribution Q_1^* of W and the conditional distribution Q_2^* of Y, given A, W. We note that $D^*(g^*, Q^*)$ can also be represented as an estimating function for ψ since $D^*(g^*, Q^*) = D^*(\Psi(Q^*), g^*, Q^*)$, as we did above.

Let Q_{2n}^{*0} be an initial estimator of $Q_{20}^*(A,W) = P_0^*(Y = 1 \mid A, W)$ according to a particular working model $\mathcal{Q}^w$ for Q_{20}^*: for example,

$$Q_{2n}^{*0} = \arg \max_{Q_2^* \in \mathcal{Q}^w} \sum_{i=1}^n q_0 \log Q_2^*(A_{1i}, W_{1i}) + \frac{\bar{q}_0(M_{1i})}{J} \sum_{j=1}^J \log(1 - Q_2^*(A_{2i}^j, W_{2i}^j)),$$

or the logistic regression based estimator Q_{n,q_0}^* using an intercept adjustment in terms of $\log q_0/(1-q_0)$ presented in Section 2 of the accompanying technical report.

Given a model $\mathcal{G}$ for g_0^*, let g_n^* be the corresponding weighted MLE:

$$g_n^* = \arg \max_{g \in \mathcal{G}} \sum_{i=1}^n q_0 \log g(A_{1i} \mid W_{1i}) + \frac{\bar{q}_0(M_{1i})}{J} \sum_{j=1}^J \log g(A_{2i}^j \mid W_{2i}^j).$$

Similarly, let Q_{1n}^* be the nonparametric weighted MLE:

$$Q_{1n}^* = \arg \max_{Q_1} \sum_{i=1}^n q_0 \log dQ_1(W_{1i}) + \frac{\bar{q}_0(M_{1i})}{J} \sum_{j=1}^J \log dQ_1(W_{2i}^j),$$

where the maximum is over all discrete distributions which put mass on W_{1i} and W_{2i}, $i = 1, \ldots, n$. It follows that Q_{1n}^* is a discrete distribution which puts mass q_0/n on W_{1i}, $i = 1, \ldots, n$, and puts mass $\bar{q}_0(M_{1i}))/(nJ)$ on W_{2i}^j, $j = 1, \ldots, j$, $i = 1, \ldots, n$.

Given any Q^*, g^*, let $\{Q_{2g^*}^*(\epsilon) : \epsilon\}$ be a model through Q_2^* at $\epsilon = 0$ and satisfying that the span of its score at $\epsilon = 0$ includes the component $D_1^*(g^*, Q^*)$ of the efficient influence curve of Ψ under i.i.d. sampling from P_{Q^*,g^*}^*. For example,

$$\frac{d}{d\epsilon} \log \left\{ Q_{2g^*}^*(\epsilon)^Y (1 - Q_{2g^*}^*(\epsilon))^{1-Y} \right\} \Big|_{\epsilon=0} = D_1^*(g^*, Q^*).$$

This can be achieved with the following fluctuation function of Q_2^*:

$$\text{logit} Q_{2g^*}^*(\epsilon) = \text{logit} Q_2^* + \epsilon Z(g^*),$$

where

$$Z(g^*) \equiv \left\{ \frac{I(A=1)}{g^*(1 \mid M,W)} - \frac{I(A=0)}{g^*(0 \mid M,W)} \right\}.$$

Given the estimator g_n^* of g_0^*, consider the fluctuation function $\{Q_{2ng_n^*}^{*0}(\epsilon) : \epsilon\}$ and let ϵ_n^0 be its weighted MLE:

$$\epsilon_n^0 = \arg\max_{\epsilon} \sum_{i=1}^{n} q_0 \log Q_{2ng_n^*}^{*0}(\epsilon)(A_{1i}, W_{1i}) + \frac{\bar{q}_0(M_{1i})}{J} \sum_{j=1}^{J} \log(1 - Q_{2ng_n^*}^{*0}(\epsilon)(A_{2i}^j, W_{2i}^j)),$$

which can be computed with standard logistic regression software.

The first step targeted MLE is now defined as

$$(g_n^*, Q_{1n}^*, Q_{2n}^{*1} = (g_n^*, Q_{1n}^*, Q_{2n}^0(\epsilon_n^0)).$$

The k-th step targeted MLE is given by $(g_n^*, Q_{1n}^*, Q_{2n}^{*k} = Q_{2n}^{*k-1}(\epsilon_n^{k-1}))$, where, for $k = 0, \ldots$

$$\epsilon_n^k = \arg\max_{\epsilon} \sum_{i=1}^{n} q_0 \log Q_{2ng_n^*}^{*k}(\epsilon)(A_{1i}, W_{1i}) + \frac{\bar{q}_0(M_{1i})}{J} \sum_{j=1}^{J} \log(1 - Q_{2ng_n^*}^{*k}(\epsilon)(A_{2i}^j, W_{2i}^j)).$$

The corresponding k-th step targeted MLE of ψ_0 is defined as $\psi_n^k = \Psi(Q_n^{*k}) \equiv \Psi(Q_{1n}^*, Q_{2n}^{*k})$. In this particular application, it follows that convergence occurs in one step so that $\psi_n = \Psi(Q_n^{*1})$.

The case-control weighted double robust estimating function for case control data is given by:

$$\begin{aligned} D_{q_0}(g^*, Q^*)(O) &= q_0 D^*(g^*, Q^*)(M_1, W_1, A_1, 1) \\ &\quad + \frac{\bar{q}_0(M_1)}{J} \sum_{j=1}^{J} D^*(g^*, Q^*)(M_1, W_2^j, A_2^j, 0), \end{aligned}$$

and the targeted MLE (g_n^*, Q_n^*) solves

$$0 = \sum_{i=1}^{n} D_{q_0}(g_n^*, Q_n^*)(O_i).$$

Statistical inference for ψ_n can be derived from the corresponding estimating equation $0 = \sum_{i=1}^{n} D(\psi_n, g_n^*, Q_n^*)(O_i)$ solved by the targeted MLE $\psi_n = \Psi(Q_n^*)$.

2.7 Double robust locally efficient targeted MLE of treatment specific mean, causal relative risk and odds ratio for case control design I.

Let $\tilde{Q}_n^*$ be defined as a standard logistic regression fit ignoring the case control sampling. Subsequently, we map this into our estimator Q_{n,q_0}^* of Q_0^* by adding the intercept $\log c(q_0)$ to the log odds of $\tilde{Q}_n^*$.

We now construct an ϵ-fluctuation $Q_{n,q_0}^*(\epsilon)$ through the corresponding logistic regression fit $Q_{n,q_0}^*(Y \mid A, W)$ satisfying

$$\frac{d}{d\epsilon} \log Q_{n,q_0}^*(\epsilon) = D^*(Q_{n,q_0}^*, g_n^*),$$

where $D^*(Q^*, g^*)$ is the efficient influence curve of the bivariate parameter $(\Psi(Q^*)(0), \Psi(Q^*)(1))$ (i.e. EY_0, EY_1). This can be done by adding a two dimensional extension $\epsilon(I(A=1)/g_n^*(1 \mid W), I(A=0)/g_n^*(0 \mid W))$ to the log odds of the logistic regression fit Q_{n,q_0}^*.

Let

$$\epsilon_n = \arg\max_{\epsilon} \sum_i q_0 \log Q^*(W_{1i}, A_{1i}) + (1-q_0)\frac{1}{J}\sum_j \log(1 - Q^*(W_{2i}^j, A_{2i}^j))$$

be the case control weighted maximum likelihood estimator of ϵ, which can be fitted with standard logistic regression software again. The one-step targeted MLE of Q_0^* is now defined as $Q_n^* \equiv Q_{n,q_0}^*(\epsilon_n)$.

Since the update of the MLE Q_{n,q_0}^* only depends on g_n^* which does not change, it follows that this one-step targeted MLE Q_n^* already solves the case-control weighted efficient influence curve estimating equation:

$$\begin{aligned} 0 &= \sum_i q_0 D^*(Q_n^*, g_n^*)(W_{1i}, A_{1i}, 1) + (1-q_0)\frac{1}{J}\sum_j D^*(Q_n^*, g_n^*)(W_{2i}^j, A_{2i}^j, 0) \\ &\equiv \sum_i D_{q_0}(Q_n^*, g_n^*)(O_i), \end{aligned}$$

so that the generally prescribed iteration for targeted MLE is not needed.

The resulting targeted maximum likelihood estimator $\Psi(Q_n^*) = E_{Q_{W,n}^*} Q_n^*(a, W)$, with $Q_{W,n}^* = q_0 Q_{W_1,n}^* + (1-q_0) Q_{W_2,n}^*$ being the case control weighted empirical distribution of the covariate vector W, solves now the double robust estimating equation $0 = \sum_i D_{q_0}(Q_n^*, g_n^*, \Psi(Q_n^*))(O_i)$ (where we now use the estimating function representation of $D_{q_0}^*$), and is therefore a double robust estimator in the sense that it is consistent and asymptotically linear if either Q_n^* is consistent or g_n^* is consistent.

The same statistical properties are now established for the corresponding causal relative risks and odds ratios, where one uses that $Q_n^* = Q_{n,q_0}^*(\epsilon_n)$, just like Q_{n,q_0}^*, equals q_0 times a bounded estimator $Q_n^\#$ so that the standard error of this double robust targeted MLE is proportional to q_0 (divided by $\sqrt{n}$).

3 Case-control weighting of efficient procedure yields an efficient procedure for both case-control designs I and II.

In this section we state and show the remarkable nice result that assigning the case-control weights to the case-control sample and then applying an efficient procedure developed for prospective sampling actually yields an efficient procedure. These results are presented and derived for both case-control designs.

3.1 Case-control weighted mapping maps gradients into gradients.

Consider a target parameter $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ at P^* in model $\mathcal{M}^*$. The class of all regular asymptotically linear estimators of $\Psi^*(P^*)$ at P^* can be characterized by their influence curves, and their influence curves constitute the set of gradients of the pathwise derivative of Ψ^* at P^* given a rich class of parametric fluctuations through P^*. In particular, an estimator is asymptotically efficient at P^* if and only if its influence curve equals the canonical gradient, that is, the unique gradient which is also an element of the tangent space generated by the scores of the class of parametric fluctuations. As a consequence of these general and powerful results an estimation problem is essentially characterized by the class of gradients and the canonical gradient. In particular, the class of gradients yields the class of wished estimating functions to construct double robust locally efficient estimators (van der Laan and Robins (2002)) and the canonical gradient provides the fundamental ingredient of the double robust locally efficient targeted maximum likelihood estimator.

This motivates us to identify the class of gradients, and, in particular, the canonical gradient, of the parameter Ψ^* in the case-control sampling model $\mathcal{M} = \{P(P^*, \eta) : P^* \in \mathcal{M}^*, \eta\}$ implied by the model $\mathcal{M}^*$ for the probability distribution P^* of interest and possible specification of dependence as identified by the η parameter, assuming that this parameter Ψ^* can be identified from case-control sampling.

The following theorem establishes that the case-control weighting does provide a mapping from the set of all gradients of the parameter $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ at P^* in model $\mathcal{M}^*$ into a set of gradients of $\Psi : \mathcal{M} \to \mathbb{R}^d$ defined as $\Psi(P(P^*, \eta)) = \Psi^*(P^*)$ at $P(P^*, \eta)$ in model $\mathcal{M} = \{P(P^*, \eta) : P^* \in \mathcal{M}^*, \eta\}$ for parameters Ψ^* which are identifiable from $P(P^*, \eta)$ (e.g. by being a function of q_0 or $\bar{q}_0(M)$). Since the class of all gradients of a parameter defined on a model represents the class of all possible influence curves of regular asymptotically linear estimators (see e.g, Bickel et al. (1993)), this result teaches us that the case-control weighting does map any estimation procedure developed for ψ_0^* based on prospective data into a corresponding estimation procedure based on case-control data, at least, from an asymptotic point of view.

In addition, since the case-control weighted mapping is 1-1, it also teaches us that it maps into a very rich set of estimation procedures for case-control data, if not all estimation procedures of interest: Indeed, we will show in the next subsections that the case-control weighted gradient mapping maps, in particular, into the optimal canonical gradient/efficient influence curve.

If the parameter of interest $\Psi^*(P^*)$ is only identified from $P = P(P^*, \eta)$ if q_0 and (for matched case-control designs) $\bar{q}_0$ is known, then one needs to define the parameter as a parameter indexed by the known q_0 and $\bar{q}_0(M)$: $\Psi^* = \Psi^*_{q_0}$.

We start with providing a useful definition of a gradient of a pathwise derivative.

Definition 2 *We define a gradient of pathwise derivative of the parameter $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ at P^* in model $\mathcal{M}^*$ as a function $D^*(P^*)$ satisfying for each of the submodels $\{P^*_{S^*}(\epsilon) : \epsilon\} \subset \mathcal{M}^*$ through P^* at $\epsilon = 0$ with score S^* at $\epsilon = 0$ (within the class of submodels through P^* specified)*

$$\frac{d}{d\epsilon}\Psi^*(P^*_{S^*}(\epsilon))\Big|_{\epsilon=0} = -\frac{d}{d\epsilon}P^*D(P^*_{S^*}(\epsilon))\Big|_{\epsilon=0}.$$

Consider a parameter $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ which is identified in model $\mathcal{M} = \{P = P(P^*, \eta) : P^* \in \mathcal{M}^*, \eta\}$, and corresponding parameter $\Psi : \mathcal{M} \to \mathbb{R}^d$ defined as $\Psi(P(P^*, \eta)) = \Psi^*(P^*)$.

By the same definition of a gradient above, a gradient of the pathwise derivative of the parameter $\Psi : \mathcal{M} \to \mathbb{R}^d$ at $P = P(P^*, \eta)$ in model $\mathcal{M}$ is defined as a function $D(P^*, \eta)$ of O satisfying for each sub-model $\{P(P^*_{S^*}(\epsilon), \eta_{S_1}(\epsilon)) : \epsilon\} \subset \mathcal{M}$ implied by a submodel $\{P^*_{S^*}(\epsilon) : \epsilon\}$ through P^* and a nuisance submodel $\{\eta_{S_1}(\epsilon) : \epsilon\}$ through η indexed by S_1,

$$\Psi^*(P^*_{S^*}(\epsilon))|_{\epsilon=0} = -\frac{d}{d\epsilon}PD(P^*_{S^*}(\epsilon), \eta_{S_1}(\epsilon))\Big|_{\epsilon=0}.$$

Given this definition of a gradient we obtain the following theorem.

Theorem 3 *Given a $P^* \in \mathcal{M}^*$, a class of sub-models $\{P^*_{S^*}(\epsilon) : \epsilon\} \subset \mathcal{M}^*$ through P^* at $\epsilon = 0$ indexed by S^*, with score S^*, we have for each of these submodels*

$$\frac{d}{d\epsilon} P D_{q_0}(P^*_{S^*}(\epsilon))\Big|_{\epsilon=0} = \frac{d}{d\epsilon} P^* D^*(P^*_{S^*}(\epsilon))\Big|_{\epsilon=0}, \tag{2}$$

where it is assumed that the left and right derivative exist.

By (2) it follows that any gradient $D^*(P^*)$ of $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ at $P^* \in \mathcal{M}^*$ is mapped into a gradient $D_{q_0}(P^*)$ of $\Psi : \mathcal{M} \to \mathbb{R}^d$ at $P = P(P^*, \eta)$ (for each η) in the model $\mathcal{M}$.

This last statement is an immediate consequence of (2) and the fact that $D_{q_0}(P^*)$ does only depend on $P = P(P^*, \eta)$ through P^* (and thus not through η), so that the derivatives along nuisance models $\{\eta(\epsilon) : \epsilon\}$ are zero, as required.

We now note that under extremely weak regularity conditions, the above definition of a gradient $D^*(P^*)$ of the pathwise derivative exactly agrees with the definition of a gradient of the pathwise derivative of $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ in efficiency theory (e.g., Bickel et al. (1993)), and similarly for Ψ. Namely, the equivalence follows if the second equality below holds (the first follows since $D^*(P^*) \in L_0^2(P^*)$): for the function $P^* \to D^*(P^*) \in L_0^2(P^*)$ and each submodel $\{P^*(\epsilon) : \epsilon\}$ (for each $P^* \in \mathcal{M}^*$) we have

$$\begin{aligned} \frac{1}{\epsilon} P^* D^*(P^*(\epsilon)) &= -\frac{1}{\epsilon} \int D^*(P^*(\epsilon)) \frac{dP^*(\epsilon) - dP^*}{dP^*(\epsilon)} dP^*(\epsilon) \\ &= -P^* D^*(P^*) S(P^*) + o(1), \end{aligned}$$

where $S(P^*)$ is the score $\frac{d}{d\epsilon} \log dP^*(\epsilon)/dP^*\Big|_{\epsilon=0}$ of the submodel $\{P^*(\epsilon) : \epsilon\}$.

For the interested reader, the following analogue theorem states the result in terms of the gradient of the pathwise derivative as in efficiency theory. That is, it provides the regularity condition under which we have that if $D^*(P^*)$ is a gradient of Ψ^* at P^*, then $D_{q_0}(P^*)$ is a gradient of the path-wise derivative of Ψ at $P(P^*, \eta)$.

Theorem 4 *Assume $\Psi : \mathcal{M} \to \mathbb{R}^d$ satisfies $\Psi(P(P^*, \eta)) = \Psi^*(P^*)$ for all $P^* \in \mathcal{M}^*$ and η.*

Assume $P^* \to D^*(P^*)$ is a gradient of the pathwise derivative of $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ in the sense that it satisfies for each member of a class of submodels $\{P^*_{S^*}(\epsilon) : \epsilon\}$ through $P^* \in \mathcal{M}^*$ at $\epsilon = 0$ with score S^*

$$\frac{d}{d\epsilon} \Psi^*(P^*_{S^*}(\epsilon))\Big|_{\epsilon=0} = -\frac{d}{d\epsilon} P^* D^*(P^*_{S^*}(\epsilon))\Big|_{\epsilon=0},$$

and the right-hand side equals $P^*D^*(P^*)S^*$, where it is assumed the derivative on the left and right-hand side exist.

Assume $P^* \to D_{q_0}(P^*)$ satisfies for each submodel $\{P(\epsilon) = P(P^*(\epsilon), \eta(\epsilon)) : \epsilon\} \subset \mathcal{M}$ through $P(P^*, \eta)$ at $\epsilon = 0$ (implied by the class of submodels $\{P^*_{S^*}(\epsilon)\}$ and $\{\eta_{S_1}(\epsilon)\}$) with score $S(P)$ that

$$-\frac{d}{d\epsilon} PD_{q_0}(P^*(\epsilon))\Big|_{\epsilon=0} = PD_{q_0}(P^*)S(P).$$

The latter is a regularity condition since

$$\begin{aligned}\frac{1}{\epsilon} PD_{q_0}(P^*(\epsilon)) &= -\frac{1}{\epsilon}\int D_{q_0}(P^*(\epsilon))\frac{dP(\epsilon) - dP}{dP(\epsilon)} dP(\epsilon) \\ &= -PD_{q_0}(P^*)S(P) + o(1),\end{aligned}$$

where $S(P)$ is the score $\frac{d}{d\epsilon}\log dP(\epsilon)/dP\Big|_{\epsilon=0}$ of the submodel $\{P(\epsilon) : \epsilon\}$.

Then, $\Psi : \mathcal{M} \to \mathbb{R}^d$ is pathwise differentiable in the sense that for each of the submodels $\{P(\epsilon) = P(P^*(\epsilon), \eta(\epsilon)) : \epsilon\} \subset \mathcal{M}$ through $P(P^*, \eta)$ at $\epsilon = 0$ with score $S(P)$ we have

$$\frac{d}{d\epsilon}\Psi(P(\epsilon))\Big|_{\epsilon=0} = PD_{q_0}(P)S(P),$$

and $D_{q_0}(P)$ is a gradient of the pathwise derivative.

Thus, for each gradient $D^*(P^*)$ of the pathwise derivative of $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ satisfying the above mentioned regularity conditions, the corresponding $D_{q_0}(P^*)$ is a gradient of the pathwise derivative of $\Psi : \mathcal{M} \to \mathbb{R}^d$.

Proof. We have

$$\begin{aligned}\frac{\Psi(P(\epsilon)) - \Psi(P)}{\epsilon} &= \frac{\Psi^*(P^*(\epsilon)) - \Psi^*(P^*)}{\epsilon} \\ &= -\frac{d}{d\epsilon} P^*D^*(P^*(\epsilon))\Big|_{\epsilon=0} + o(1) \\ &= -\frac{d}{d\epsilon} PD_{q_0}(P^*(\epsilon))\Big|_{\epsilon=0} + o(1) \\ &= PD_{q_0}(P^*)S(P) + o(1).\end{aligned}$$

This proves that $\Psi : \mathcal{M} \to \mathbb{R}^d$ defined as $\Psi(P(P^*, \eta)) = \Psi^*(P^*)$ is pathwise differentiable at $P = P(P^*, \eta) \in \mathcal{M}$ and that $D_{q_0}(P^*)$ is a gradient of this pathwise derivative. □

Thus, the above result shows that each gradient $D^*(P^*)$ for $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ is mapped into a gradient $D_{q_0}(P^*)$ for $\Psi : \mathcal{M} = \{P(P^*, \eta) : P^* \in \mathcal{M}^*, \eta\} \to \mathbb{R}^d$ defined as $\Psi(P(P^*, \eta)) = \Psi^*(P^*)$. We note that this gradient mapping is not affected by the particular choice (i.e., model of dependence structure of case and control observations) of model $\mathcal{M} = \{P(P^*, \eta) : P^* \in \mathcal{M}^*, \eta\}$ compatible with $\mathcal{M}^*$. Thus, for example, for case-control design I, our mapping from gradients into gradients for model $\mathcal{M}$ is the same for the independence model assuming the case and controls are all independent as it is for a particular dependence model.

A particular case is that $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ is defined on a nonparametric model $\mathcal{M}^*$. In this case, there exists only one gradient for model $\mathcal{M}^*$ so that one just needs to determine the canonical gradient $D^*(P^*)$ of Ψ^* at P^* and map it into its case-control weighted version $D_{q_0}(P^*)$, which, by our results in the next section, equals the canonical gradient of Ψ at $P(P^*, \eta)$.

Remark. Since q_0 is a non-identifiable parameter for both case-control designs (so that knowledge of q_0 does not restrict the distribution of the data structure O), this implies that 1) for each gradient $D^*(P^*)$ for model $\mathcal{M}^*$, the corresponding $D_{q_0}(P^*)$ is a gradient in the model $\mathcal{M}$ *also including* the knowledge that q_0 is known (even if that knowledge was not included in $\mathcal{M}^*$), or, equivalently, the class of all gradients $\{D_h^*(P^*) : h\}$ at P^* for model $\mathcal{M}^*$ is mapped into a class $\{D_{h,q_0} : h\}$ of gradients at $P = P(P^*)$ for model $\mathcal{M}$ also including q_0 is known.

For matched case-control design II, if we define our parameter as $\Psi^*_{q_0}$, indexed by q_0 and $\bar{q}_0(M)$ (treating them as known and fixed), then the case-control weighting maps the class of all gradients of this parameter for model $\mathcal{M}^*$ into the class of gradients of this parameter for model $\mathcal{M} = \{P(P^*, \eta) : P^* \in \mathcal{M}^*, \eta\}$. If the observed data model is the same with and without the restriction that $(q_0, \bar{q}_0(M))$ is known in the model $\mathcal{M}^*$, then the canonical gradient in the model $\mathcal{M}$ will be the same as the canonical gradient of the model also including the knowledge of $(q_0, \bar{q}_0(M))$.

3.2 Independence models for case-control designs I and II to derive efficiency results.

We consider the independence model $\mathcal{M}$ so that $\mathcal{M} = \{P(P^*) : P^* \in \mathcal{M}^*\}$, where for case-control design I, we have

$$dP(P^*)(W_1, A_1, (W_2^j, A_2^j : j)) = dP^*(W_1, A_1 \mid Y = 1) \prod_{j=1}^{J} dP^*(W_2^j, A_2^j \mid Y = 0), \tag{3}$$

and, for case-control design II, we have

$$\begin{aligned} dP(P^*)(M_1, W_1, A_1, (M_1, W_2^j, A_2^j : j)) &= dP^*(M_1, W_1, A_1 \mid Y = 1) \\ &\prod_{j=1}^{J} dP^*(W_2^j, A_2^j \mid M = M_1, Y = 0). \\ &= dP_M^*(M_1) dP^*(W_1, A_1 \mid M = M_1, Y = 1) \\ &\prod_{j=1}^{J} dP^*(W_2^j, A_2^j \mid M = M_1, Y = 0). \end{aligned} \tag{4}$$

Our results immediately generalize to models $\mathcal{M}$ for which the densities of the distributions $P(P^*, \eta)$ factorize as

$$dP(P^*, \eta) = dP_1(P^*) dP_2(\eta),$$

where $dP_1(P^*)$ is given by the independence likelihood (3) or (4), and P^* and η are variation independent. This follows from the fact that such models the tangent space contains the tangent space of the independence model, and our proof of the wished result is based on showing that the case-control weighted efficient influence curve is a member of the tangent space and thereby equals the efficient influence curve for the model $\mathcal{M}$.

Our results in this section show that the case-control weighting of the canonical gradient for the prospective sampling model $\mathcal{M}^*$ yields the canonical gradient for the parameter of interest Ψ based on case-control sampling model $\mathcal{M}$. Our results rely on the assumption that (the typically very large/semiparametric) $\mathcal{M}^*$ corresponds with (i.e., equals the intersection of) separate models for $P_0^*(W, A \mid Y = \delta)$ for $\delta \in \{0, 1\}$ for case-control design I, and that $\mathcal{M}^*$ corresponds with (i.e., equals the intersection of) separate models for $P_0^*(W, A \mid Y = \delta, M = m)$ for $\delta \in \{0, 1\}$ and m varying over the support of the matching variable M.

As a consequence of our results, our proposed case-control weighted targeted maximum likelihood estimator for variable importance and causal effect

parameters, involving selecting estimators of Q_0^* and g_0^*, under appropriate regularity conditions guaranteeing the wished convergence of the standardized estimator to a normal limit distribution, is efficient if both of these estimators are consistent, and remains consistent if one of these estimators is consistent.

We note that the working-model to obtain the initial model based maximum likelihood estimators in our double robust targeted maximum likelihood estimator is obtained by modeling the factors of

$$dP^*(W, A, Y) = dP^*(W)dP^*(A \mid W)dP^*(Y \mid A, W),$$

which does thus not correspond with separate models for $dP^*(W, A \mid Y = \delta)$ as we "required" for the actual model $\mathcal{M}^*$ in order to make sure that the case-control weighted canonical gradient is a canonical gradient. In order to understand the rational of this discrepancy we provide the following explanation.

It happens to be that the efficient influence curve for our parameter of interest Ψ for an underlying model $\mathcal{M}^*$ identified by separate models for $P(W, A \mid Y = \delta)$ has a double robust representation in terms of Q_0^* and g_0^*, while it does not have a double robust representation w.r.t. to say $P(W, A \mid Y)$ or factors thereof. To fully exploit this double robust representation of the efficient influence curve of our parameter of interest, one should base estimation of the unknowns parameters of the efficient influence curve on the latter representation, and that is why we proposed our particular double robust locally efficient targeted maximum likelihood estimators.

Alternatively, we could use a targeted maximum likelihood estimator based on initial estimators based on working models for $P(W, A \mid Y = \delta)$, $\delta \in \{0, 1\}$: in this manner we would obtain generalized locally efficient double robust estimators where the double robustness is stated in terms of the models for Q_0^* and g_0^* *implied by* the models for $P(W, A \mid Y = \delta)$.

3.3 Case-control weighting of canonical gradient yields canonical gradient: Case Control Design I.

Firstly, we present the theorem for case-control design I.

Theorem 5 *Consider case-control design I. Assume that the model $\mathcal{M}^*$ allows independent variation of $P^*(W, A \mid Y = 1)$ and $P^*(W, A \mid Y = 0)$.*

Let $D^*(P^*)$ be the canonical gradient of the pathwise derivative $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ at $P^* \in \mathcal{M}^*$, let $\mathcal{M} = \{P(P^*) : P^* \in \mathcal{M}^*\}$ be the independence model defined by (3), and let $\Psi : \mathcal{M} \to \mathbb{R}^d$ satisfy $\Psi(P(P^*)) = \Psi^*(P^*)$ for all

$P^* \in \mathcal{M}^*$. Assume the regularity conditions for $P^* \to D^*(P^*)$ of Theorem 4 apply so that it follows that Ψ is pathwise differentiable at P^* and $D_{q_0}(P^*)$ is a gradient of this pathwise derivative.

We have that $D_{q_0}(P^*)$ is the canonical gradient of the pathwise derivative of $\Psi : \mathcal{M} \to \mathbb{R}^d$.

We already knew that, if we set $D^*(P^*)$ equal to the canonical gradient (or any other gradient) of $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$, then its case-control weighted version $D_{q_0}(P^*)$ is a gradient of $\Psi : \mathcal{M} \to \mathbb{R}^d$. The surprising and important extra result is that this $D_{q_0}(P^*)$ actually equals the canonical gradient. That is, for case-control design I, the case-control weighted gradient mapping does not only map gradients into gradients, it also maps the optimal canonical gradient for model $\mathcal{M}^*$ into the optimal canonical gradient for the observed data model $\mathcal{M}$ for case-control data.

Remark regarding q_0 known in model $\mathcal{M}^*$. Since q_0 is a non-identifiable parameter based on case-control sampling (design I), assuming q_0 is known in model $\mathcal{M}^*$ puts no restriction on the observed data model $\mathcal{M}$. As a consequence, the efficient influence curve for the parameter $\Psi : \mathcal{M} \to \mathbb{R}^d$ is the same for the model $\mathcal{M}^*$ in which this quantity is known as it is in the model in which this quantity is unknown.

3.4 Example of efficient method for case-control design II based on stratified efficient method for case-control design I.

Before we present our general analogue result for case-control design II, it is helpful to consider an example for case-control design II. Consider the data structure $O^* = (M, W, A, Y) \sim P_0^*$ and let $\mathcal{M}^*$ be a nonparametric model. Consider case-control design II, in which our observed data $O = ((M_1, W_1, A_1), ((W_2^j, A_2^j) : j = 1, \ldots, J))$. Suppose we wish to estimate $\psi_0^* = E_0^* Y_1 = E_0^* E_0^*(Y \mid A = 1, M, W)$ and that $q_0(\delta \mid m) = \delta P_0^*(Y = 1 \mid M = m) + (1 - \delta) P_0^*(Y = 0 \mid M = m)$ is known. Recall that the efficient influence curve for this parameter $\Psi^* : \mathcal{M}^* \to \mathbb{R}$ in model $\mathcal{M}^*$ at P^* is given by $D^*(Q^*, g^*) - \psi^* = I(A = 1)/g^*(1 \mid M, W)(Y - Q^*(M, W, A)) + Q^*(M, W, 1) - \psi^*$.

Consider the following general approach for estimation of ψ_0^* based on data generated by a case-control design II:

- Apply the case-control weighted targeted MLE for case-control design I to the subsample $\{i : M_{1i} = m\}$ to estimate the conditional version $\psi_0^*(m) = E^*(Y_1 \mid M = m)$ of the parameter ψ_0^*. Thus this corresponds with weighting the cases with $q_0(1 \mid m) = P_0^*(Y = 1 \mid M = m)$ and the controls with $q_0(0 \mid m) = P_0^*(Y = 0 \mid M = m)$ and applying the standard prospective targeted MLE based on an initial estimator of $Q_0^*(m, a, w) = P_0^*(Y = 1 \mid m, a, w)$ and $g_0^*(a \mid m, w) = P_0^*(A = a \mid M = m, W = w)$. By our results for case-control design I, we know that this estimator yields a double robust locally efficient estimator of $\psi_0(m)$.

  This case-control weighted targeted maximum likelihood estimator of $\psi_0(m)$ based on the subsample $\{i : M_{1i} = m\}$ solves the m-specific case-control weighted efficient influence curve equation $0 = P_n D^*_{m,q_0}(Q_n^*, g_n^*) - \Psi^*(Q_n^*)(m)$ and can thus be represented as

$$\psi_n(m) = \frac{\sum_i I(M_{1i} = m) D_{m,q_0}(Q_n^*, g_n^*)(O_i)}{\sum_i I(M_{1i} = m)}, \tag{5}$$

  where

$$\begin{aligned} &D_{m,q_0}(Q^*, g^*)(O) = \\ &q_0(1 \mid m) \left\{ \tfrac{I(A_1=1)}{g_0^*(1|m,W_1)}(1 - Q^*(m, W_1, 1)) + Q^*(m, W_1, 1) \right\} \\ &+ \tfrac{q_0(0|m)}{J} \left\{ \tfrac{I(A_2^j=1)}{g^*(1|m,W_2^j)}(0 - Q^*(m, W_2^j, A_2^j, 1)) + Q^*(m, W_2^j, A_2^j, 1) \right\}. \end{aligned}$$

  The rational behind the consistency of this estimator $\psi_n(m)$ follows directly from the identity

$$E(Y_1 \mid M = m) = \frac{E_0 D_{m,q_0}(Q_0^*, g_0^*)(O) I(M_1 = m)}{P_0(M_1 = m)}.$$

- Now, note that

$$P_0^*(M = m) = P_0(M_1 = m) \frac{q_0}{q_0(1 \mid m)}.$$

  Thus, one maps $\psi_n(m)$ into an estimator of ψ_0 by averaging it w.r.t. to $q_0/q_0(1 \mid M_{1i}) P_n(M_1 = m)$:

$$\begin{aligned} \psi_n &= \sum_m \left\{ \frac{1}{n} \sum_{i=1}^n I(M_{1i} = m) \frac{q_0}{q_0(1 \mid M_{1i})} \right\} \psi_n(m) \\ &= \frac{1}{n} \sum_{i=1}^n \sum_m \frac{q_0}{q_0(1 \mid m)} I(M_{1i} = m) D_{m,q_0}(Q_n^*, g_n^*)(O_i), \end{aligned}$$

where we used (5).

Again, the rational of this estimator of ψ_0 follows immediately from the following derivation:

$$\begin{aligned}
&E_0 \textstyle\sum_m \frac{q_0}{q_0(1|m)} I(M_1 = m) D_{m,q_0}(Q_0^*, g_0^*) \\
&= E_0 \tfrac{q_0}{q_0(1|M_1)} D_{M_1,q_0}(Q_0^*, g_0^*) \\
&= E_0 \tfrac{q_0}{q_0(1|M_1)} \left\{ q_0(1 \mid M_1) D^*(M_1, W_1, A_1, 1) + \textstyle\sum_j \frac{q_0(0|M_1)}{J} D^*(M_1, W_2^j, A_2^j, 0) \right\} \\
&= E_0 q_0 D^*(M_1, W_1, A_1, 1) + \tfrac{\bar{q}_0(M_1)}{J} \textstyle\sum_j D^*(M_1, W_2^j, A_2^j, 0) \\
&= E_0^* Y_1,
\end{aligned}$$

where we suppressed the dependence of $D^* = D^*(Q^*, g^*)$ on Q^*, g^*.

- We conclude that this estimator ψ_n of ψ_0^* corresponds with solving our proposed case-control weighted efficient influence curve equation $P_n D_{q_0,\bar{q}_0} - \psi = 0$, where

$$D_{q_0,\bar{q}_0}(O) = q_0 D^*(M_1, W_1, A_1, 1) + \frac{\bar{q}_0(M_1)}{J} \sum_j D^*(M_1, W_2^j, A_2^j, 0).$$

We conclude that this general approach for estimation of ψ_0^* of applying the case-control weighted targeted MLE $\psi_n(m)$ of case-control design I to the sub-sample $\{i : M_{1i} = m\}$ to estimate the analogue $\psi_0^*(m)$ of the parameter of interest ψ_0^* (i.e., the same function but now applied to the conditional $P_0^*(\cdot \mid M = m)$), and subsequently averaging $\psi_n(m)$ w.r.t. $q_0/q_0(1 \mid m) P_n(M_1 = m)$, corresponds with using our for case-control design II proposed case-control weighting $D_{q_0,\bar{q}_0}$ of the efficient influence curve D^* for model $\mathcal{M}^*$. This suggests that $D_{q_0,\bar{q}_0}$ is indeed also, just as we showed for case-control design I, the efficient influence curve. Our results below confirm this.

3.5 Case-control weighting of canonical gradient yields canonical gradient: Matched Case Control Design.

For case-control design II, we establish the same result.

Theorem 6 *Consider case-control design II. In this theorem we use the notation:* $D_{q_0,\bar{q}_0}(P^*) = q_0 D^*(P^*)(M_1, W_1, A_1, 1) + \frac{\bar{q}_0(M_1)}{J} \sum_j D^*(P^*)(M_1, W_2^j, A_2^j, 0)$.

Assume that the model $\mathcal{M}^*$ allows independent variation of $P^*(W, A \mid Y = \delta, M = m)$ for $\delta \in \{0, 1\}$ and possible outcomes m of M under P_0^*.

Let $D^*(P^*)$ be the canonical gradient of the pathwise derivative $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ at $P^* \in \mathcal{M}^*$, let $\mathcal{M} = \{P(P^*) : P^* \in \mathcal{M}^*\}$ be the independence model

defined by (4), and let $\Psi : \mathcal{M} \to \mathbb{R}^d$ *satisfy* $\Psi(P(P^*)) = \Psi^*(P^*)$ *for all* $P^* \in \mathcal{M}^*$.

Assume the regularity conditions for $P^* \to D^*(P^*)$ *of Theorem 4 apply so that it follows that* Ψ *is pathwise differentiable and* $D_{q_0,\bar{q}_0}(P^*)$ *is a gradient of this pathwise derivative at* $P(P^*) \in \mathcal{M}$.

Then, $D_{q_0,\bar{q}_0}(P^*)$ *is the canonical gradient of the pathwise derivative of* $\Psi : \mathcal{M} \to \mathbb{R}^d$.

3.6 Selecting the efficient influence curve of unrestricted target parameter.

In order to define an identifiable parameter $\Psi(P(P^*)) = \Psi^*(P^*)$ of the case-control data generating distribution, one often needs to define Ψ^* as indexed by the known q_0 and possibly $\bar{q}_0$ parameters. We denote such a parameter with $\Psi^*_{q_0} : \mathcal{M}^* \to \mathbb{R}^d$ to stress its dependence on these known fixed quantities. Our results above for case-control designs I and II above prove that if $D^*(P^*)$ is the canonical gradient of $\Psi^*_{q_0}$ at P^*, then the case-control weighted $D_{q_0}(P^*)$ is the canonical gradient of $\Psi : \mathcal{M} \to \mathbb{R}^d$, where $\Psi(P(P^*)) = \Psi^*_{q_0}(P^*)$ for all $P^* \in \mathcal{M}$. The following theorem shows that one can typically replace $D^*(P^*)$ by the canonical gradient of the path-wise derivative of the unrestricted $\Psi^*(P^*) = \Psi_{q(P^*)}(P^*)$.

Theorem 7 *Consider the two pathwise differentiable parameters* $\Psi^*_{r_0} : \mathcal{M}^* \to \mathbb{R}^d$ *indexed by a fixed* $r_0 = r(P^*_0)$ *(e..g, representing* q_0 *and* $\bar{q}_0$*), and a corresponding parameter* $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ *defined as* $\Psi^*(P^*) = \Psi^*_{r(P^*)}(P^*)$. *Thus,* $\Psi^*_{r_0}(P^*_0) = \Psi^*(P_0)$.

Assume that for all the sub-models $\{P^*(\epsilon) : \epsilon\}$ *for which* $\frac{d}{d\epsilon} r(P^*(\epsilon))\Big|_{\epsilon=0} = 0$, *we have*

$$\frac{d}{d\epsilon}\Psi^*(P^*(\epsilon))\Big|_{\epsilon=0} = \frac{d}{d\epsilon}\Psi^*_{r_0}(P^*(\epsilon))\Big|_{\epsilon=0}.$$

Assume that the fixed parameter r_0 *in* $\Psi^*_{r_0}$ *is locally non-identifiable at* P^* *in the model* $\mathcal{M}$ *in the sense that the tangent space at* $P(P^*) \in \mathcal{M}$ *generated by the submodels* $\{P^*(\epsilon) : \epsilon\}$ *at* P^* *for which* $\frac{d}{d\epsilon} r(P^*(\epsilon))\Big|_{\epsilon=0} = 0$ *equals the tangent space at* $P(P^*) \in \mathcal{M}$ *generated by all submodels used in definition of pathwise derivative of* $\Psi^*_{r_0} : \mathcal{M}^* \to \mathbb{R}^d$.

If the conditions of Theorem 5 or Theorem 6 apply for this choice $\Psi^*_{r_0} : \mathcal{M}^* \to \mathbb{R}^d$, *then we also have, if* $D^*(P^*)$ *is the canonical gradient of* Ψ^* *at* P^*, *then the case-control weighted* $D_{q_0}(P^*)$ *is the canonical gradient of* $\Psi : \mathcal{M} \to \mathbb{R}^d$.

Proof. This result is shown as follows. Let D^* be the canonical gradient of $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ and let D_1^* be the canonical gradient of $\Psi_{q_0}^* : \mathcal{M}^* \to \mathbb{R}^d$. As a consequence of the first assumption, we have for all scores S of all these submodels $P^*(\epsilon)$ not changing q_0 (in first order),

$$\langle D^*, S\rangle_{P^*} = \langle D_1^*, S\rangle_{P^*}.$$

So, if we restrict our class of sub-models at P^* in the definition of the path-wise derivative to these sub-models in $\mathcal{M}^*$ not varying r_0 (which globally corresponds with restricting $\mathcal{M}^*$ to all P^* with $r(P^*) = r_0$, but path-wise differentiability at P^* only depends on local thickness of model at P^*), then we have that the canonical gradient for the corresponding class of submodels for the observed data model is given by the case-control weighted $D_{q_0}(P^*)$ and the latter also equals the case-control weighted $D_{1q_0}(P^*)$. So under this restriction on the class of submodels through P^* we have equality of the two case-control weighted canonical gradients corresponding with D^* and D_1^*. Now, by using that this restriction on the class of submodels does not change the tangent space for the observed data models, and therefore does not affect the canonical gradient representation at $P(P^*)$ of the parameter Ψ in the observed data model $\mathcal{M}$. Thus this $D_{q_0}(P^*)$, which equals $D_{1q_0}(P^*)$, also equals the canonical gradient for the class of all submodels used in the actual definition of the pathwise derivative. This completes the proof of the theorem. □

Since q_0 is non-identifiable for case-control design I it follows that case-control weighting of the canonical gradient of the unrestricted parameter Ψ^* also yields the wished canonical gradient of Ψ. The same would apply for the matched case-control design, if enforcing the restriction $(q_0, q_0(1 \mid m) = P_0^*(Y = 1 \mid M = m))$ in $\mathcal{M}^*$ does not reduce the observed data tangent space, but this remains to be verified.

3.7 Proof of Theorems 5 and 6.

We already know that for both designs $D_{q_0}(P^*)$ (defined as $D_{q_0,1-q_0}(P^*)$ for design I and defined as $D_{q_0,\bar{q}_0}$ for design II) is a gradient of the pathwise derivative of Ψ at $P(P^*)$. Therefore, it remains to show that $D_{q_0}(P^*)$ is an element of the tangent space $T(P(P^*)) \subset L_0^2(P(P^*))$ defined as the closure of the linear span of the scores of each of the submodels $\{P(\epsilon) : \epsilon\}$ within the Hilbert space $L_0^2(P(P^*))$.

In the Appendix we have a separate section establishing these results for both designs, stating that if we select $D^*(P^*)$ as the canonical gradient of Ψ^* at P^* and the model $\mathcal{M}^*$ allows independent variation of $P(W, A \mid Y = \delta)$ for

Design I and independent variation of $P(W, A \mid M = m, Y = \delta)$ for Design II, then $D_{q_0}(P^*)$ is an element of the tangent space at $P(P^*)$ in the observed case-control data model $\mathcal{M}$.

Here we provide a summary of the proof for case-control design I in order to provide the reader with an understanding of these results.

Since $D^*(P^*)$ is a canonical gradient it equals a score $\frac{d}{d\epsilon} dP^*(\epsilon)/dP^*\Big|_{\epsilon=0}$ for a particular submodel $\{P^*(\epsilon) : \epsilon\}$ at $\epsilon = 0$, or it can be arbitrarily well approximated by such a sequence of scores. We first consider the case that $D^*(P^*)$ is itself a score.

The tangent space under the independence model for a nonparametric model $\mathcal{M}^*$ is an orthogonal sum of the Hilbert space $T_1(P) = \{S_1(W_1, A_1) : S_1\}$ of functions of (W_1, A_1) with mean zero, and the Hilbert space $T_2(P) = \{\sum_j S_2(W_2^j, A_2^j) : S_2\}$ with $S_2(W_2^j, A_2^j)$ having mean zero, $j = 1, \ldots, J$. For an actual model $\mathcal{M}^*$ these two Hilbert spaces are replaced by sub-spaces spanned by the scores of the allowed sub-models $\{P^*(\epsilon) : \epsilon\}$ through P^*. That is, $T_1(P)$ consists of (and is generated by) functions $\frac{d}{d\epsilon}\frac{dP^*(\epsilon)}{dP^*}(W_1, A_1 \mid Y = 1)\Big|_{\epsilon=0}$, and $T_2(P)$ consists of (and is generated by) functions $\sum_j \frac{d}{d\epsilon}\frac{dP^*(\epsilon)}{dP^*}(W_2^j, A_2^j \mid Y = 0)\Big|_{\epsilon=0}$, $j = 1, \ldots, J$. We assumed that the marginal distributions $P^*(W, A \mid Y = 1)$ and $P^*(W, A \mid Y = 0)$ are independently varied by these submodels, so that indeed the tangent space is an orthogonal sum of $T_1(P)$ and $T_2(P)$.

For notational convenience, we introduce the notation $\epsilon_0 = 0$. Let $D^*(P^*) = \frac{d}{d\epsilon_0}\frac{dP^*(\epsilon_0)}{dP^*}(W, A, Y)$ be a score. Since q_0 is non-identifiable, we can assume that $p^*(\epsilon)(Y = 1) = q_0$ for all ϵ. It follows that

$$\begin{aligned}
q_0 D^*(P^*)(W_1, A_1, 1) &= q_0 \frac{1}{p^*(W_1, A_1, 1)} \frac{d}{d\epsilon_0} p^*(\epsilon_0)(W_1, A_1, 1) \\
&= q_0 \frac{1}{p^*(W_1, A_1 \mid Y = 1) q_0} \frac{d}{d\epsilon_0} p^*(\epsilon_0)(W_1, A_1 \mid Y = 1) q_0 \\
&= q_0 \frac{1}{p^*(W_1, A_1 \mid Y = 1)} \frac{d}{d\epsilon_0} p^*(\epsilon_0)(W_1, A_1 \mid Y = 1) \\
&\in T_1(P^*),
\end{aligned}$$

since the latter term equals q_0 times a score of $P(\epsilon)(W_1, A_1)$ at $\epsilon = 0$ (which in particular has mean zero).

Again, using that $P^*(\epsilon)(Y=0) = 1-q_0$ for all ϵ,

$$\begin{aligned}
&(1-q_0)D^*(P^*)(W_2^j, A_2^j, 0) = (1-q_0)\tfrac{1}{p^*(W_2^j, A_2^j, 0)}\tfrac{d}{d\epsilon_0}p^*(\epsilon_0)(W_2^j, A_2^j, 0)\\
&= (1-q_0)\tfrac{1}{p^*(W_2^j, A_2^j \mid Y=0)(1-q_0)}\tfrac{d}{d\epsilon_0}p^*(\epsilon_0)(W_2^j, A_2^j \mid Y=0)p^*(\epsilon)(Y=0)\\
&= (1-q_0)\tfrac{1}{p^*(W_2^j, A_2^j \mid Y=0)}\tfrac{d}{d\epsilon_0}p^*(\epsilon_0)(W_2^j, A_2^j \mid Y=0)\\
&\equiv (1-q_0)S_2(W_2^j, A_2^j),
\end{aligned}$$

where the latter term equals is $1-q_0$ times a score of $P(\epsilon)(W_2^j, A_2^j)$ at $\epsilon = 0$ (which, in particular, has mean zero). It follows that

$$\frac{(1-q_0)}{J}\sum_j D^*(P^*)(W_2^j, A_2^j, 0) = \frac{1-q_0}{J}\sum_j S_2(W_2^j, A_2^j) \in T_2(P(P^*)).$$

This proves that for case-control design I, if $D^*(P^*)$ is a score, then

$$D_{q_0}(P^*)(O) = q_0 D^*(P^*)(W_1, A_1) + \frac{1-q_0}{J}\sum_j D^*(P^*)(W_2^j, A_2^j)$$

is a score itself, and thus an element of the tangent space $T(P)$.

Suppose now that $D^*(P^*) = \lim_{m\to\infty} D_m^*(P^*) \in T^*(P^*)$, where $D_m(P^*) \in L_0^2(P^*)$ is a score. Then, for each m, we have $D_{mq_0}(P^*) \in L_0^2(P(P^*))$ is a score. To show that $D_{q_0}(P^*) \in L_0^2(P^*)$ is a score requires thus that the case-control mapping $D^* \to D_{q_0}$, as a mapping from $L_0^2(P^*)$ into $L_0^2(P(P^*))$ is continuous. This is trivially established. This proves that indeed $D_{q_0}(P^*)$ is an element of the tangent space $T(P(P^*))$. This completes the proof for case-control design I.

The proof for case-control design II is more delicate and provided in detail in the Appendix.

4 Summary, discussion and extensions.

We provide a generic approach for locally efficient estimation such as targeted maximum likelihood estimation of any parameter based on matched and unmatched case-control designs, which relies on specification of one or two non-identifiable parameters/scalars q_0 and, for matched case-control designs, $q_0(1 \mid m) = P_0^*(Y=1 \mid M=m)$.

These non-identifiable parameters could be known or they could be set in a sensitivity analysis, for example, in the case that these parameters are known to be contained in a particular interval. In the Appendix below we illustrate

how to handle the case that q_0 is replaced by a user supplied estimator based on an independent data set, and a standard error of this estimate of the true prevalence probability is provided.

Our approach is remarkably simple since it only requires weighting the cases by q_0 and the controls by $1 - q_0$ or $\bar{q}_0(M_1)$ and then applying a method developed for prospective sampling. Moreover, our approach has the remarkable convenient feature that applying the case-control weighting to an optimal method for the prospective sample results in an optimal method for independent and matched case-control designs.

We also showed how the case-control weighting for matched case-control designs corresponds with applying the case-control weighting for the standard unmatched case-control design for each sub-sample defined by a category for the matching variable to obtain the analogue conditional parameter, conditional on the matching variable category, and subsequently averaging these results over the matching variable categories to get the wished marginal parameter. This helps us to understand that our somewhat strange looking weights for the control observations in a matched case-control study are actually just as sensible as the much easier to understand weights for standard case-control designs.

In our accompanying technical report we worked out the case-control weighted targeted maximum likelihood estimators in a number of important applications involving estimation of variable importance and causal effect parameters. In addition, in our accompanying technical report we showed for both types of case-control designs how standard maximum likelihood logistic regression fits can be adjusted by using these known quantities to estimate conditional probabilities $P_0^*(Y = 1 \mid A, W)$ with a standard error which is proportional to q_0 divided by the square root of the sample size, so that the acquired precision results in stable estimators of such challenging parameters as relative risk and odds-ratios at $q_0 \approx 0$.

We believe that in many applications the marginal population proportion of cases, q_0, could be known, at least within close approximation, but it does require an effort to understand the target population the cases are sampled from. The literature supporting the use of q_0 in case-control studies goes back more than 50 years (See Cornfield (1951), Cornfield (1956)). Even 25 years ago, Greenland (1981) noted that "improvements in disease surveillance have produced more reliable estimates of disease incidence in many populations." Another relevant publication discussing the use of q_0 in case-control analysis is Benichou and Wacholder (1994).

In matched case-control studies in which one uses a matching variable with a large number of categories, then the value of the population proportion of

cases within each matching category might not be known. In that case, if the number of matching categories is large, a sensitivity analysis would likely be too cumbersome. On the other hand, even for such matched case-control samples, using the case-control weighting for design I might already provide an important bias reduction so that our methods only relying on q_0 will likely still provide a useful set of tools. Off course, this would require some validation that ignoring the matching does not cause severe bias.

During the design of a case-control study, we recommend to keep in mind that knowing these population proportion of cases for each matching category make the convenient and double robust efficient estimation of any causal effect and variable importance parameter possible (through the methods presented here) without restrictive assumptions such as the no-interaction assumption and parametric model form for conditional logistic regression models. This insight might help and motivate people to design case-control studies in which the required case-control weights are known or approximately known so that a sensitivity analysis is possible.

In addition, we note that the binary Y conditioned upon in the case-control sampling does not need to be an outcome of interest. For example, the random variable of interest might be a right-censored data structure $O^* = (W, A, \tilde{T} = \min(T, C))$, with T survival, C censoring, W covariates and A treatment, and in the case-control sampling we might condition upon a person having been observed to fail or not by time τ: $Y = I(\tilde{T} \leq \tau)$. In such an application the parameter of interest might be the causal effect of A on T.

To summarize, , by knowing q_0, one has available more efficient and more robust (i.e., double robust) targeted maximum likelihood estimators, targeting an identifiable parameter, and one does not have to restrict oneselves to odds-ratio parameters.

We now consider a few direct extensions and applications of our methodology.

Frequency matching. Frequency matching in case-control studies is typically defined as running a case-control design I within each strata $M = m$. In this case one can estimate any causal parameter $\psi_0(m)$ of the conditional distribution of O^*, given $M = m$, by assigning weights $q_0(1 \mid m)$ to the cases and $q_0(0 \mid m)/J$ to the corresponding J controls. Thus our methods for case-control design I can be applied to each strata $M = m$. In particular, this yields a locally efficient double robust targeted maximum likelihood estimator of $\psi_0(m)$ for each m. In order to estimate the marginal parameter ψ_0 one would need an estimate of the marginal distribution of M, which cannot be

identified based on knowing $q_0(1 \mid m)$ only, so that other knowledge will be needed such as the marginal population distribution of M. Either way, one can always estimate causal parameters such as $E(Y_a \mid M = m)$ for each m or the corresponding variable importance measure. If the number of categories of the matching variable is large, then a sensible strategy for estimation of $\psi_0(m)$ is to assume a model $\psi_0(m) = f(m \mid \beta_0)$ and obtain a pooled locally efficient targeted maximum likelihood of β_0 based on all observations.

Pair matching. Pair matching in case-control studies is typically described as, for each matching category, sample a case and a set of controls. So this description agrees with frequency matching except that the number of categories can be very large. Therefore, we should now always assume a model $\psi_0(m) = f(m \mid \beta_0)$ and obtain a pooled locally efficient targeted maximum likelihood of β_0 based on all observations.

Without the knowledge of $q_0(1 \mid m)$, one would use conditional logistic regression models, and, as noted in Jewell (2006) page 258, these methods do not allow estimation of the association of M with Y, while if one knows the population proportion $q_0(1 \mid m)$ we can estimate every parameter of the population distribution, conditional on $M = m$.

Counter matching. Finally, another type of matching in case-control studies is called counter-matching, which involves sampling a control with an exposure (maximally) different from the exposure of the case. Formally, we can define this sampling scheme as follows. The observation $O = ((M_1, Z_1), (M_2, Z_2))$ on each experimental unit is generated as 1) sample (M_1, Z_1) from the conditional distribution of (M, Z), given $Y = 1$, and 2) sample (M_2, Z_2) from the conditional distribution of (M, Z), given $M = m^*(M_1)$ and $Y = 0$, where $m^*(m)$ maps a particular outcome m into a counter-match $m^*(m)$ in the outcome space for M. Similarly, this is defined for the case that one samples J controls counter-matched to the case. The population distribution of interest is the distribution P_0^* of $O^* = (M, Z, Y)$ and we are concerned with estimation of a particular parameter ψ_0^* of this distribution P_0^* based on a counter-matched case-control sample $O_1, \ldots, O_n$. In this case, given that $D^*(M, Z, Y)$ satisfies $P_0^* D^* = 0$, we have

$$E_0 D_{q_0, \bar{q}_0^*}(O) = 0,$$

where the case-control weighted version of D^* is defined as

$$D_{q_0, \bar{q}_0}(O) = q_0 D^*(M_1, Z_1, 1) + \bar{q}_0^*(M) D^*(m^*(M_1), Z_2, 0),$$

with

$$\bar{q}_0^*(m) = (1 - q_0)\frac{P_0^*(M = m^*(m) \mid Y = 0)}{P_0^*(M = m \mid Y = 1)}.$$

Note that if $m^*(m) = m$ is the identity function, then indeed $\bar{q}_0^* = \bar{q}_0$. The non-identifiable component of the control-weight $\bar{q}_0^*$ is $P_0^*(M = m^*(m), Y = 0)$, or, assuming q_0 is known, $P_0^*(M = m^*(m) \mid Y = 0)$, while the denominator $P_0^*(M = m \mid Y = 1) = P_0(M_1 = m)$ can be empirically estimated. Since in many applications the control observations are relatively easily accessible, one might use a separate sample of controls to estimate these proportions $P_0^*(M = \cdot \mid Y = 0)$ having a certain value for the (counter-)matching variable M among the controls. So under the condition that these weights $q_0, \bar{q}_0^*$ are known (or set in a sensitivity analysis), our results in this article can be applied to counter-matched case-control designs by just replacing $\bar{q}_0$ by $\bar{q}_0^*$.

Propensity score matching design. A commonly used design is the following. One samples from the units that received treatment. For each treated unit, one finds a matched non-treated unit, where the matching is done based on a fit of the so called propensity score. The goal of this design is to create a sample in which the confounders are reasonably balanced between the treated and untreated units. This design can formally be described as follows. The random variable of interest is $O^* = (W, A, Y) \sim P_0^*$, and one is typically concerned with estimation of a causal effect such as $E_0^*\{E_0^*(Y \mid A = 1, W) - E_0^*(Y \mid A = 0, W)\}$. Let $M \equiv \Pi^*(W)$ be a summary measure of W which is supposedly an approximation of the propensity score $\Pi_0^*(W) = P_0(A = 1 \mid W)$ (e.g., estimated from external data). One samples $(M_1 = \Pi^*(W_1), W_1, Y_1)$ from the conditional distribution of (W, Y), given $A = 1$, and one samples one or more $(M_2 = \Pi^*(W_2), W_2, Y_2)$ from the conditional distribution of (W, Y), given $M = M_1$and $A = 0$.

One now wishes to use n i.i.d. observations on the observed experimental unit $O = ((W_1, Y_1), (W_{2j}, Y_{2j} : j))$ representing a treated unit and one or more propensity score matched untreated units to estimate the causal parameter of interest.

Notice that we can immediately apply the methodology presented in this article by defining the Y as the A and the matching variable M is playing the role of $\Pi^*(W)$. As a consequence, one can use any method developed for sampling from (W, A, Y) by using our "case control" weights $q_0 = P_0^*(A = 1)$ for the treated units, and $\bar{q}_0(W) = q_0 \frac{P_0^*(A=0|M)}{P_0^*(A=1|M)}$ for the untreated units. Thus, to correct for the biased sampling one will need to know the actual true treatment mechanism/propensity score $P_0^*(A = 1 \mid W)$. Thus, under

the assumption that this propensity score is known or can be estimated based on an external data source, one can apply any method for estimation of the wished causal effect for standard sampling by applying these weights to the treated and untreated units. Off course, for the sake of statistical inference and model selection (say, based on cross-validation) one should respect the fact that the independent and identically distributed observations are $O_1, \ldots, O_n$, and not the treated and untreated units.

General biased sampling. Finally, we like to discuss the implications of the proposed optimal case-control weighting for general biased sampling models with known probabilities for the conditioning events, where optimal refers to the fact that the case-control weighting maps an efficient procedure for an unbiased sample into an efficient procedure for the biased sample. The following generalization of our method for case-control design I applies to general biased sampling. Consider a particular target probability distribution P_0^* representing the unbiased sampling distribution and its corresponding random variable $O^* \sim P_0^*$. Suppose now that the outcome space for the random variable O^* is partitioned by a union of events $\mathcal{A}_j$, $j = 1, \ldots, J$: i.e. $Pr(O^* \in \cup_j \mathcal{A}_j) = 1$ and the sets $\mathcal{A}_j$ are pairwise disjoint. Let the experimental unit for the observed data be $(O_1, \ldots, O_J)$, where $O_j \sim O^* \mid O^* \in \mathcal{A}_j$ is a draw from the conditional distribution, given $O^* \in \mathcal{A}_j$, $j = 1, \ldots, J$. For simplicity, we enforced here equal number of draws, but this can be generalized to having different number of draws from each conditional distribution. Let $q_0(j) = P_0^*(O^* \in \mathcal{A}_j) \in (0, 1)$ and suppose these probabilities are known. Weighting observation O_j with $q_0(j)$ for $j = 1, \ldots, J$, and applying a method developed for the unbiased sample will yield valid estimators. We also conjecture that under appropriate similar conditions as we assumed for case-control sampling, this weighting will be optimal in the sense that assigning these weights to an efficient estimation procedure for i.i.d. samples of P_0^* will yield an efficient estimation procedure based on the biased sampling model. Given our interpretation of case-control weighting for matched case-control sampling in terms of case-control weighting for standard case-control studies conditional on the matching category, we suggest that weighting for matched case-control sampling can be generalized to matched biased sampling in general (say matched on a draw M_1 from the first biased sampling distribution).

Another commonly employed study is a case-control sample nested within a cohort. In addition, it is then common that one collects additional information on the case-control sample relative to the information collected in the original cohort sample. Our results are not covering this important problem for which

a rich literature exist (see e.g., Robins et al. (1994)).

Appendix: Incorporating variability/uncertainty in the user supplied prevalence probability q_0.

In this section we wish to illustrate that our general case-control weighted estimation methodology directly generalizes to the case that q_0 is replaced by an estimate $\hat{q}$ (based on an independent sample) with a user supplied standard error σ. For the sake of illustration, consider the independent case-control design and let $D_{q_0}(O \mid \psi) = q_0 D(W_1, A_1, 1 \mid \psi) + (1 - q_0)/J \sum_j D(W_0^j, A_0^j, 0 \mid \psi)$ be a case-control weighted estimating function applied to an estimating function $D(O^* \mid \psi)$ for the parameter of interest $\psi_0 = \Psi(P_0^*)$ of the target distribution P_0^*. Let the case-control weighted estimator $\hat{\Psi}(q_0, P_n)$ be defined as a solution of the estimating equation

$$0 = P_n D_{q_0}(O \mid \psi) = \frac{1}{n} \sum_{i=1}^{n} D_{q_0}(O_i \mid \psi),$$

where P_n denotes the empirical distribution.

The case-control weighted estimator ψ_n based on $\hat{q}$ of ψ_0 can now be represented as $\hat{\Psi}(\hat{q}, P_n)$. Under regularity conditions, the estimator $\hat{\Psi}(q_0, P_n)$ (as consider in our article) using the true prevalence probability q_0 is asymptotically linear with influence curve $IC_0 = -\frac{d}{d\psi_0} P_0 D_{q_0}(\psi_0)^{-1} D_{q_0}(\psi_0)$, using short-hand notation. The actual estimator $\hat{\Psi}(\hat{q}, P_n)$ can now be decomposed as

$$\begin{aligned} \hat{\Psi}(\hat{q}, P_n) - \psi_0 &= \hat{\Psi}(\hat{q}, P_n) - \hat{\Psi}(\hat{q}, P_0) + \hat{\Psi}(\hat{q}, P_0) - \hat{\Psi}(q_0, P_0) \\ &\approx \hat{\Psi}(q_0, P_n) - \hat{\Psi}(q_0, P_0) + \hat{\Psi}(\hat{q}, P_0) - \hat{\Psi}(q_0, P_0), \end{aligned}$$

where the approximation involves a second order term of $\hat{q} - q_0$ and $P_n - P_0$. The first difference equals $(P_n - P_0)IC_0 + o_P(1/\sqrt{n})$ and is thus asymptotically normally distribution with mean zero and covariance matrix $\Sigma_0 = E_0 IC_0 IC_0^\top$. The second difference is independent of this first asymptotically normal term and, by the delta-method, can be approximated by $\hat{q} - q_0$ times the gradient a_0 of $q \to \hat{\Psi}(q, P_0)$:

$$\hat{\Psi}(\hat{q}, P_0) - \hat{\Psi}(q_0, P_0) = (\hat{q} - q_0) \frac{d}{dq_0} \hat{\Psi}(q_0, P_0) = (\hat{q} - q_0) a_0.$$

Thus, this term behaves as a normally distributed vector with mean zero and variance elements $\sigma^2 a_0$, where $a_0 = \frac{d}{dq_0}\hat{\Psi}(q_0, P_0)$. We can conclude that our standardized estimator $\sqrt{n}(\hat{\Psi}(\hat{q}, P_n) - \psi_0)$ converges in distribution to

$$N(0, \Sigma + \sigma^2 a_0 a_0^\top),$$

where $\Sigma = E_0 IC_0(O) IC_0^\top(O)$ is the covariance matrix of the normal limit distribution of the estimator $\hat{\Psi}(q_0, P_n)$ based on the known prevalence probability.

In general, this general template shows that we can incorporate the standard error σ of a user supplied estimate $\hat{q}$ by simply adding the matrix $\sigma^2 a_0 a_0^\top$ to the covariance matrix of our case-control weighted estimator $\hat{\Psi}(\hat{q}, P_n)$ we would use if $\hat{q}$ is treated as known, where a_0 is the gradient of $q \to \hat{\Psi}(q, P_0)$ at q_0.

For the sake of concreteness, we will now provide an expression of the gradient a_0 of the derivative of $q \to \hat{\Psi}(q, P_0)$ at $q = q_0$ in the above setting. Note that $\hat{\Psi}(q, P_0)$ is defined as the solution in ψ of $H_0(q, \psi) = P_0 D_{q_0}(\psi) = 0$. By the implicit function theorem, this shows that the gradient of $q \to \hat{\Psi}(q, P_0)$ is given by:

$$\begin{aligned} a_0 &= -\frac{d}{d\psi_0} H_0(q_0, \psi_0)^{-1} \frac{d}{dq_0} H_0(q_0, \psi_0) \\ &= -\frac{d}{d\psi_0} H_0(q_0, \psi_0)^{-1} P_0(D_1 - D_0), \end{aligned}$$

where we defined $D_1(O) = D(1, W_1, A_1)$ and $D_0(O) = \frac{1}{J}\sum_j D(0, W_0^j, A_0^j)$. One can estimate a_0 by replacing the expectations by empirical means, and thereby construct confidence intervals and p-values based on $\Sigma_n + \sigma^2 a_n a_n^\top$, where Σ_n is an estimator of the covariance matrix Σ_0 and a_n is the estimator of a_0.

Appendix: Extension to case-control incidence density sampling.

An alternative commonly employed case-control sampling design involves regular case-control sampling from a population at risk at time t, where the outcome is now defined at time t, across various time points t (see e.g., Rothman and Greenland (1998)). Such designs can be carried out at only a few discrete time points or they could evolve in continuous time.

For example, one might sample breast cancer cases and controls in year 2000 among the population at risk of breast cancer, and one would repeat

such a case-control sample at years 2001 and 2002. Note that the outcome is now different depending on the year one samples, since being a case in the case-control sample at year (e.g.) 2000 requires being diagnosed with breast cancer in year 2000. Another type of example would be to sample one or more controls at the time a case occurs among the subjects at risk right before the case occurred.

One issue with this kind of case-control sampling is that the sampling population might change over time due to an influx of new subjects over time, so that the change in sampling population over time cannot only be modeled by censoring and the occurrence of failures within a well defined target population at the first time point. Alternatively, one defines a target population at the first time point and one samples cases and controls at time t among the subjects in this target population that are still at risk right before time t (i.e., the subjects that have not failed or been censored, yet), thereby ignoring any possibly influx of subjects over time.

We now wish to discuss some possible applications of our case-control weighting methodology to these types of case-control sampling designs. Firstly, the most straightforward and direct application is to treat the case-control sample at time t as a separate case-control sample and immediately apply our case-control weighting to estimate any parameter of the population distribution one samples from at time t. Of course, this requires a large enough case-control sample at each time point t so that these t-specific parameters are estimated at a reasonable precision. Note also that the knowledge of the case-control weights now requires knowing the marginal probability of being a case for the sampling population at time t, at each of the sampling times t. If one is willing to assume that these t-specific parameters (e.g. causal effect of a treatment on outcome) follow a parametric trend in t, then one can pool all the t-specific estimates to obtain a smoother estimation procedure that might result in significant gains in variance. For example, maybe it is appropriate to believe that the population is stationary in time t, that is, somehow the influx of new subjects and loss of existing subjects due to censoring or failure balances out so that the sampling population at time t does not change over time. In that case, one might assume that the t-specific parameters are constant in t.

We now wish to consider how we might generally apply pooling across time while using our case-control weighting to handle such incidence density sampling designs. Here, we will focus on a single target population so that one is concerned with estimation of a single well defined parameter of a target population of interest.

Consider the case that the outcome of interest is a time till event T. For

notational convenience, we will assume that T is discrete on time points $t = 0, 1, \ldots, \tau$. Suppose that in a prospective sample one would observe $O^* = (\tilde{T} = \min(C, T), \Delta = I(\tilde{T} = T), \bar{X}(\tilde{T}))$, where C is a right-censoring time and $\bar{X}(s)$ denotes the history up till time s of the time dependent process $t \to X(t)$: $\bar{X}(s) = (X(u) : u \leq s)$, where $X(t)$ includes the indicator $dY(t) \equiv I(T = t)$ of the failure time event at time t. Let P_0^* denote the probability distribution of this right-censored data structure O^*. Suppose that the parameter of interest is $\Psi(P_0^*)$ which will typically represent a parameter of the full data distribution of X such as a causal effect of a treatment A assigned at time 0 on the time till event T.

In the case that the outcome is a time till a *rare* event one might employ a so called incidence density case-control sampling design. That is, at time t, among the population at risk defined by all the individuals with $R(t) = I(\tilde{T} \geq t) = 1$, one samples a case from the conditional distribution of O^*, given $dY(t) = 1$ and $R(t) = 1$, and one samples one or more controls from the conditional distribution of O^*, given $dY(t) = 0$ and $R(t) = 1$. Note that one can replace $dY(t)$ by the observed data quantity $dY(t) = I(\tilde{T} = t, \Delta = 1)$. Let's denote the observed data structure sampled at time t, consisting of a case and one or more controls, as

$$O_t = (O_{1t}, O_{0tj},\ j = 1, \ldots, J),$$

where O_{1t} denotes the data structure on the case and O_{0tj} denotes the data structure on the j-th control. Suppose one samples $n(t)$ i.i.d observations of O_t at time t, $t = 0, \ldots, \tau$, resulting in a total sample O_{ti}, $i = 1, \ldots, n(t)$, $t = 0, \ldots, \tau$.

Let $R(t)D(t, O^*)$ be an estimating function or loss function for the prospectively sampled unit O^*, $t = 0, \ldots, \tau$. An estimating function or loss function based on sampling O^* itself can always be represented as $\sum_t R(t)D(t, \bar{O}^*(t))$, where $\bar{O}^*(t) = \bar{X}(\min(t, C))$ denotes the observed history up till time t, which is assumed to include the censoring event if it occurs before time t. Specifically, we have

$$\begin{aligned} D(O^*) &= \sum_t E(D \mid \bar{X}(\min(t, C))) - E(D \mid \bar{X}(\min(t-1, C))) \\ &= \sum_t R(t) \left\{ E(D \mid \bar{X}(\min(t, C)) - E(D \mid \bar{X}(\min(t-1, C))) \right\}, \end{aligned}$$

where $\bar{X}(\min(t, C))$ represents the history one observes up till time t, and thus it is assumed that $\bar{X}(\min(t, C))$ also includes observing the censoring event time C if C occurs before time t

The following lemma shows how the case-control weighting can be applied to this t-specific estimating function of O^* which typically represents just one

term $R(t)D(t,O^*)$ of the full estimating function $D(O^*) = \sum_t R(t)D(t,O^*)$ one would use if one would sample O^* prospectively.

Lemma 1 *Define*

$$D_{q_0}(t,O_t) \equiv q_0(t)D(t,O_{1t}) + \bar{q}_0(t)\frac{1}{J}\sum_{j=1}^{J} D(t,O_{0tj}),$$

where

$$\begin{aligned} q_0(t) &\equiv P_0^*(dY(t)=1, R(t)=1) \\ \bar{q}_0(t) &\equiv P_0^*(dY(t)=0, R(t)=1). \end{aligned}$$

We have

$$E_0 D_{q_0}(t,O_t) = E_0^* R(t)D(t,O^*).$$

In particular, if we redefine $q_0(t) = P(dY(t)=1 \mid R(t)=1)$ *and* $\bar{q}_0(t) = 1-q_0(t)$*, then*

$$E_0 D_{q_0}(t,O_t)P_0^*(R(t)=1) = E_0^* R(t)D(t,O^*).$$

If censoring is non-informative, then the weights $q_0(t) = P_0^*(dY(t)=1 \mid R(t)=1) = P_0^*(dY(t)=1 \mid T \geq t)$ reduce to the marginal hazard of T at time t. Thus, if censoring is non-informative, then this case-control weighting would require knowing the marginal failure time distribution of T.

Proof of Lemma. We have

$$\begin{aligned} E_0 D_t(O_t) &= E_0 q_0(t)D(t,1,O_1^*) + \bar{q}_0(t)\frac{1}{J}\sum_{j=1}^{J} D(t,0,O_{0j}^*) \\ &= \int D(t,1,O_1^*)q_0(t)P_0^*(O^* \mid dY(t)=1, R(t)=1) \\ &\quad + \frac{1}{J}\sum_{j=1}^{J}\int D(t,0,O^*)\bar{q}_0(t)P_0^*(O^* \mid dY(t)=0, R(t)=1) \\ &= \int_{O^*} D(t,1,O^*)P_0^*(O^*, dY(t)=1, R(t)=1) \\ &\quad + \frac{1}{J}\sum_{j=1}^{J}\int_{O^*} D(t,0,O^*)P_0^*(O^*, dY(t)=0, R(t)=1) \\ &= \int_{O^*,R(t)} R(t)D(t,1,O^*)P_0^*(O^*, dY(t)=1, R(t)) \\ &\quad + \int_{O^*,R(t)} R(t)D(t,0,O^*)P_0^*(O^*, dY(t)=0, R(t)) \\ &= \int_{O^*,dY(t),R(t)} R(t)D(t,dY(t),O^*)P_0^*(O^*, dY(t), R(t)). \end{aligned}$$

This proves the lemma. □

Even though one only applies the t-specific component $R(t)D(t,O^*)$ of the full estimating function to the case, the following lemma shows that one can often use the control observation sampled at time t for the later time point estimating functions without any need for weighting or coupling them to the case sampled at time t.

Lemma 2 *Assume $E_0(D(s,O^*) \mid R(s)=1)=0$ for all s. Given a t, for $s>t$, we have for the control observations O_{0t}*

$$E_0R(s)D(s,O_{0t})=0$$

Proof. We have for $s>t$

$$\begin{aligned}
&E_0R(s)D(s,O_{0t}) = E_0(R(s)D(s,O^*) \mid R(t)=1,dY(t)=0)\\
&= E_0(E_0(R(s)D(s,O^*) \mid R(s),R(t)=1,dY(t)=0) \mid R(t)=1,dY(t)=0)\\
&= E_0(P_0(R(s)=1 \mid R(t)=1,dY(t)=0)\\
&\times E_0(R(s)D(s,O^*) \mid R(s)=1,R(t)=1,dY(t)=0) \mid R(t)=1,dY(t)=0)\\
&= P_0(R(s)=1 \mid R(t)=1,dY(t)=0)\\
&\times E_0(E_0(R(s)D(s,O^*) \mid R(s)=1) \mid R(t)=1,dY(t)=0)\\
&= P_0(R(s)=1 \mid R(t)=1,dY(t)=0)E_0(D(s,O^*) \mid R(s)=1)\\
&= 0.\square
\end{aligned}$$

Thus, given an estimating function $D(O^* \mid \psi)=\sum_t R(t)D(t,O^* \mid \psi)$ for the parameter ψ_0^* based on sampling from P_0^*, an estimating equation for the total sample from the actual biased sampling data generating distribution P_0 can now be constructed as:

$$\begin{aligned}
0 \;=\; &\sum_t \sum_{i=1}^{n(t)} q_0(t)D(t,O_{1ti}) + \bar{q}_0(t)\frac{1}{J}\sum_{j=1}^{J} D(t,O_{0tji} \mid \psi)\\
&+\sum_t \sum_{i=1}^{n(t)} \sum_{s>t} \frac{1}{J}\sum_j R(s)D(s,O_{0sji} \mid \psi).
\end{aligned}$$

The last term represents the non-coupled contributions of the control observations at time points after the time point t at which the control unit was sampled. Here $q_0(t)=P(dY(t)=1 \mid R(t)=1)$ and $\bar{q}_0(t)=P(dY(t)=0 \mid R(t)=0)=1-q_0(t)$. If the estimating function is indexed by nuisance parameters, then these need to be estimated.

Not conditioning on being at risk. In the above form of incidence density sampling, sampling a case corresponds with conditioning on a subject being at risk and being a true case at time t (i.e., a failure at time t). In the following lemma we employ the same design but in which sampling a case corresponds with only conditioning on having an observed event at time t, and thus not conditioning on being at risk at time t. The advantage of this type of design is that one can now case-control weight the complete estimating function $D(O^*) = \sum_t R(t)D(t, O^*)$ one would use in prospective/unbiased sampling of O^*. The lemma provides the correct case control weighting and it is now a direct corollary of our case-control weighting results established in this article.

Lemma 3 *Let $dY(t) = I(\tilde{T} = t, \Delta = 1)$. Let O_{1t} be a draw from conditional distribution of O^*, given $dY(t) = 1$, and let O_{0tj} be i.i.d draws from the conditional distribution of O^*, given $dY(t) = 0$, $j = 1, \ldots, J$. Let $O_t = (O_{1t}, (O_{0tj} : j))$ be the total observation consisting of a case and J controls.*

Given a function $O^* \rightarrow D(O^*)$, define

$$D_{q_0}(t, O_t) \equiv q_0(t)D(O_{1t}) + \bar{q}_0(t)\frac{1}{J}\sum_{j=1}^{J} D(O_{0tj}),$$

where

$$\begin{aligned} q_0(t) &\equiv P_0(dY(t) = 1) \\ \bar{q}_0(t) &\equiv P_0(dY(t) = 0). \end{aligned}$$

We have

$$E_0 D_{q_0}(t, O_t) = E_0^* D(O^*).$$

Thus, given an estimating function $D(O^* \mid \psi) = \sum_t R(t)D(t, O^* \mid \psi)$ for the parameter ψ_0^* based on sampling from P_0^*, an estimating equation for the total sample from the actual biased sampling data generating distribution P_0 can now be constructed as

$$0 = \sum_t \sum_{i=1}^{n(t)} q_0(t)D(O_{1ti} \mid \psi) + \bar{q}_0(t)\frac{1}{J}\sum_{j=1}^{J} D(O_{0tji} \mid \psi),$$

$q_0(t) = P(dY(t) = 1)$ and $\bar{q}_0(t) = P(dY(t) = 0) = 1 - q_0(t)$. If the estimating function is indexed by nuisance parameters, then these need to be estimated.

If there is censoring, then, even if censoring is independent, $q_0(t)$ is also a function of the censoring mechanism for the prospective data structure O^*,

which might be viewed as a disadvantage of such weights. Again, we note that the case-control weighting requires knowing these marginal incidence probabilities $q_0(t)$ across all time points t to which one applies the case control sampling.

Matched case-control incidence density sampling. Since the weights of our lemmas are directly implied by our case-control weights used in this article, it is also clear how we can generalize the above lemmas to matched case-control incidence density sampling in which one matches the controls by also conditioning on a matching variable being equal to the matching variable of the case.

An example: Estimation of conditional hazards based on incidence density case-control sampling. Consider a target population of individuals, and suppose that the data structure O^* on a sampled individual consists of baseline covariates W, a treatment variable A, and a right-censored survival time T, so that $O^* = (W, A, \tilde{T} = \min(T, C), \Delta = I(\tilde{T} = T)$. Suppose we are concerned with estimation of an intensity $E(dY(t) \mid \bar{F}(t), A, W)$ of the counting process $Y(t) = I(\tilde{T} \leq t, \Delta = 1)$ w.r.t to history $\bar{F}(t), A, W$, where $\bar{F}(t)$ represents the failure and censoring history up till time t. If censoring is conditionally independent of T given A, W, then this intensity equals $I(\tilde{T} \geq t)E(I(T \in dt) \mid T \geq t, A, W)$, that is, it equals the conditional hazard of T, given A, W. It is common to assume a Cox-proportional hazards or logistic regression model, depending on T being continuous (or finely discrete) or discrete. For the sake of illustration, let's consider a parametric model $\alpha_\beta(t \mid \bar{F}(t), A, W)$ for this intensity $E(dY(t) \mid \bar{F}(t), A, W)$ indexed by a finite dimensional parameter β. Under sampling from O^* it is known how to construct a good estimator of β_0. In particular one can apply maximum likelihood estimation where the likelihood for a single observation O^* is given by $\prod_t P_\beta(dY(t) \mid W, A, \bar{F}(t))$ in which

$$P_\beta(dY(t) \mid W, A, \bar{F}(t)) = \\ \alpha_\beta(t \mid W, A, \bar{F}(t))^{dY(t)}(1 - \alpha_\beta(dY(t) \mid W, A, \bar{F}(t))^{1-dY(t)}$$

represents the Bernoulli likelihood corresponding with the model α_β. Let $R(t)D_\beta(t, dY(t), W, A, \bar{F}(t))$ be defined as t-specific term $R(t) \log P_\beta(dY(t) \mid W, A, \bar{F}(t))$ of the log-likelihood $\sum_t R(t) \log P_\beta(dY(t) \mid W, A, \bar{F}(t))$ of O^*.

Consider now a case-control incidence density sampling design in which at time t one samples a case from the conditional distribution of O^*, given that $dY(t) = 1, R(t) = 1$, and one or more controls from the conditional

distribution of O^*, given that $dY(t) = 0$, $R(t) = 1$. Let $O_t = (O_{1t}, (O_{0tj} : j))$ denote the coupled case and control observations. As above, let $n(t)$ denote the number of such observations one samples at time t: O_{ti}, $i = 1, \ldots, n(t)$. For notational convenience, let's assume that one samples a single control for each case: i.e., $J = 1$. As case-control weighted log-likelihood, augmented with the control observation contributions for later time points, we obtain:

$$\begin{aligned} L_n(\beta) &= \sum_t \sum_{i=1}^{n(t)} q_0(t) D_\beta(t, 1, W_{1ti}, A_{1ti}, \bar{F}_{1ti}(t)) \\ &\quad + \bar{q}_0(t) D_\beta(t, 0, W_{0ti}, A_{0ti}, \bar{F}_{0ti}(t)) \\ &\quad + \sum_t \sum_{i=1}^{n(t)} \sum_{s>t} R(s) D_\beta(s, 0, W_{0ti}, A_{0ti}, \bar{F}_{0ti}(s)). \end{aligned}$$

The time-specific case-control weights are $q_0(t) = P(\tilde{T} = t, \Delta = 1 \mid \tilde{T} \geq t)$ and $\bar{q}_0(t) = 1 - q_0(t)$. If censoring C is independent of T, then $q_0(t) = P(T = t \mid T \geq t)$ is the marginal hazard of T. One can now apply standard maximum likelihood estimation to this log-likelihood, which can be carried out with standard software.

This case-control weighted log-likelihood can also be written down for a semi-parametric Cox-proportional hazards model, and, again, the corresponding maximum likelihood estimator can be found by using standard maximum likelihood estimation software developed for fitting a Cox-model based on prospective sampling.

Finally, in an anologue fashion we can now obtain case-control weighted targeted maximum likelihood estimators of particular parameters of this intensity such as marginal causal effects of A on T, but we reserve this for future work.

Appendix: Tangent space results proving case-control weighted canonical gradient of prospective sampling model equals canonical gradient.

Our results in this section show that the case-control weighted canonical gradient for the prospective sampling model $\mathcal{M}^*$ yields the canonical gradient for the parameter of interest Ψ in the actual case-control sampling model. These results rely on the following assumption. The (typically very large/semiparametric) model $\mathcal{M}^*$ corresponds with (i.e., equals the intersection of) separate models for $P_0^*(W, A \mid Y = \delta)$ for $\delta \in \{0, 1\}$ for case-control

design I, and, for case-control design II, $\mathcal{M}^*$ corresponds with (i.e., equals the intersection of) separate models for $P_0^*(W, A \mid Y = \delta, M = m)$ for $\delta \in \{0, 1\}$ and m varying over the support of the matching variable M. As a consequence of this canonical gradient representation our proposed case-control weighted targeted maximum likelihood estimator, involving selecting estimators of Q_0^* and g_0^*, under appropriate regularity conditions guaranteeing the wished convergence to a normal limit distribution, is efficient if both of these estimators are consistent, and remains consistent if one of these estimators is consistent.

The results are stated in an incremental fashion thereby building up the proof of the final wished result. As a consequence, most stated results do not require a proof but can be straightforwardly verified.

Tangent space for case-control design I. We start out with presenting the tangent space for case-control design I.

Theorem 8 *(***Tangent space for case-control design I***) Consider case-control design I and the independence model $\mathcal{M}$ described by (3),*

$$dP(P^*)(O) = P^*(W_1, A_1 \mid Y = 1) \prod_j P^*(W_2^j, A_2^j \mid Y = 0),$$

and let $T^*(P^*)$ denote the tangent space at P^* in model $\mathcal{M}^*$. The tangent space at $P(P^*)$ in model $\mathcal{M}$ is given by

$$T_I(P^*) = \\ \left\{ S^*(W_1, A_1, 1) - E^*(S^* \mid Y = 1) + \textstyle\sum_j \{S^*(W_2^j, A_2^j, 0) - E^*(S^* \mid Y = 0)\} \right\},$$

where S^* varies across $T^*(P^*)$.

Since this tangent space is expressed in terms of the tangent space of the underlying model $\mathcal{M}^*$ we now need to understand the tangent space of $\mathcal{M}^*$. The following theorem fully characterizes this tangent space for models $\mathcal{M}^*$ described by separate models for $P(W, A \mid Y = \delta)$ for $\delta \in \{0, 1\}$.

Theorem 9 *(***Tangent space for underlying model $\mathcal{M}^*$***) Consider the data structure $O^* = (W, A, Y)$ and model $\mathcal{M}^*$ for its probability distribution. We make the following assumption on $\mathcal{M}^*$: Let $\mathcal{M}^* = \cap_\delta \mathcal{M}^*(\delta)$, where $\mathcal{M}^*(\delta)$ is a model for $P_0^*(W, A \mid Y = \delta)$ indexed by (possibly infinite dimensional) parameter $\theta(\delta)$, for each $\delta \in \{0, 1\}$, and assume that $\theta(\delta)$ for different choices of δ are variation independent parameters.*

If the marginal distribution $q_0(\delta) = P(Y = \delta)$ of Y is known in model $\mathcal{M}^*$, then, we can represent $T^*(P^*)$ as

$$T^*(P^*) = \sum_{\delta} T^*_{\delta}(P^*), \tag{6}$$

where the latter sum-space is an orthogonal sum, and $T^*_{\delta}(P^*)$ denotes the tangent space generated by $\theta(\delta)$, which can be represented as

$$T^*_{\delta}(P^*) = \{I(Y = \delta)\,(S^*(W, A, \delta) - E(S^* \mid Y)) : S^* \in T^*(P^*)\}\,.$$

If $q_0(\delta)$ is unknown and modelled, then

$$T^*(P^*) = L^2_0(P^*_Y) \oplus \sum_{\delta} T^*_{\delta}(P^*), \tag{7}$$

where $L^2_0(P^*_Y)$ is the Hilbert space of functions of Y with mean zero and finite variance w.r.t. P^*. We also note that for a $S^* \in L^2_0(P^*)$, the projection of S^* on $T^*_{\delta}(P^*)$ is given by

$$\Pi(S^* \mid T^*_{\delta}(P^*)) = I(Y = \delta)\,(S^*(W, A, \delta) - E(S^* \mid Y))\,,$$

and the projection of S^* onto $T^*(P^*)$ described by the orthogonal decomposition (7) is given by

$$S^* = E(S^* \mid Y) + \sum_{\delta} \Pi(S^* \mid T^*_{\delta}(P^*)).$$

Tangent space for case-control design II. We now present the tangent space for matched case-control design II.

Theorem 10 *(***Tangent space for case-control design II***) Consider case-control design II and the independence model $\mathcal{M}$ described by (4),*

$$dP(P^*)(O) = P^*(M_1)P^*(A_1, W_1 \mid Y = 1, M_1) \prod_j P^*(A^j_2, W^j_2 \mid Y = 0, M_1),$$

and let $T^*(P^*)$ denote the tangent space at P^* in model $\mathcal{M}^*$. The tangent space at $P(P^*)$ in model $\mathcal{M}$ is given by

$$T_{II}(P^*) = L^2_0(M_1) \oplus \\ \left\{S^*(Z_1, 1) - E^*(S^* \mid M = M_1, Y = 1) + \textstyle\sum_j \{S^*(Z^j_2, 0) - E^*(S^* \mid M = M_1, Y = 0)\}\right\},$$

where S^* varies across $T^*(P^*)$, $Z_1 = (M_1, W_1, A_1)$ and $Z^j_2 = (M_1, W^j_2, A^j_2)$.

Since this tangent space is characterized in terms of the underlying tangent space $T^*(P^*)$ for model $\mathcal{M}^*$ we now fully characterize the latter tangent space for models $\mathcal{M}^*$ described by separate models for $P^*(W, A \mid M = m, Y = \delta)$ for the different values of m and δ.

Theorem 11 *(***Tangent space for model $\mathcal{M}^*$ including matching variable***) We make the following assumption on the model $\mathcal{M}^*$: Suppose that $\mathcal{M}^* = \cap_{m,\delta}\mathcal{M}^*(m,\delta)$, where $\mathcal{M}^*(m,\delta)$ is a model for $P_0^*(W, A \mid M = m, Y = \delta)$ indexed by (e.g., infinite dimensional) parameter $\theta(m,\delta)$, for each $\delta \in \{0,1\}$ and possible outcome m for M, and it is assumed that $\theta(m,\delta)$ are variation independent parameters.*

If $q_0(\delta \mid m) = P(Y = \delta \mid M = m)$ is known and the marginal distribution of M is unspecified in model $\mathcal{M}^*$, then, we can represent $T^*(P^*)$ as

$$T^*(P^*) = L_0^2(M) \oplus \sum_{m,\delta} T^*_{m,\delta}(P^*), \tag{8}$$

where the latter sum-space is an orthogonal sum, and $T^*_{m,\delta}(P^*)$ denotes the tangent space generated by $\theta(m,\delta)$, which can be represented as

$$T^*_{m,\delta}(P^*) = \{I(M = m, Y = \delta)\,(S^*(m, W, A, \delta) - E(S^* \mid M, Y)) : S^* \in T^*(P^*)\}\,.$$

If the conditional distribution $q_0(\delta \mid m)$ of Y, given M, is unknown and modeled, then

$$T^*(P^*) = L_0^2(P^*_M) \oplus T^*(q_0) \oplus \sum_{m,\delta} T^*_{m,\delta}(P^*), \tag{9}$$

where $T^*(q_0)$ denotes the tangent space generated by the scores of the parameters of $q_0(\delta \mid m)$. We also note that for a $S^* \in L_0^2(P^*)$, the projection onto $T^*_{m,\delta}(P^*)$ is given by

$$\Pi(S^* \mid T^*_{m,\delta}(P^*)) = I(M = m, Y = \delta)\,(S^*(m, W, A, \delta) - E(S^* \mid M, Y))\,,$$

and, under the assumption that $q_0(\delta \mid m)$ is unspecified, the projection of S^* onto $T^*(P^*)$ described by the orthogonal decomposition (9) is given by

$$S^* = E(S^* \mid M) + \{E(S^* \mid Y, M) - E(S^* \mid M)\} + \sum_{m,\delta} \Pi(S^* \mid T^*_{m,\delta}(P^*)).$$

Special score for case-control design I. We will later show that the case-control weighted canonical gradient is in the tangent space $T_I(P^*)$ by selecting a special choice $S^* \in T^*(P^*)$ defined in the next result. The following result shows that this special choice is indeed a member of $T^*(P^*)$.

Result 1 *Let $O^* = (W, A, Y) \sim P_0^* \in \mathcal{M}^*$ and assume that the tangent space $T^*(P^*)$ at $P^* \in \mathcal{M}^*$ is given by orthogonal decomposition (7). Given a $D^* \in T^*(P^*)$, we have*

$$\begin{aligned} S^*(W, A, Y) &= q_0(Y) \{D^*(W, A, Y) - E^*(D^* \mid Y)\} \\ &\in T^*(P^*). \end{aligned}$$

The same applies if $q_0(0)$ is replaced by $q_0(0)/J$.

Proof. Firstly, we note that for each δ, $\Pi(D^* \mid T_\delta(P^*)) \in T^*(P^*)$, and by linearity of the space $T_\delta(P^*)$ (i.e., closure under multiplication by scalar) we have that $q_0(\delta)\Pi(D^* \mid T_\delta^*(P^*)) \in T^*(P^*)$. By linearity of $T^*(P^*)$, it follows thus that

$$\begin{aligned} &\textstyle\sum_\delta q_0(\delta)\Pi(D^* \mid T_\delta^*(P^*)) \\ &= \textstyle\sum_\delta q_0(\delta) I(Y = \delta) \left(D^*(W, A, \delta) - E^*(D^* \mid Y)\right) \\ &= q_0(Y) \left(D^*(W, A, Y) - E^*(D^* \mid Y)\right) \\ &= S^*(W, A, Y) \\ &\in T^*(P^*). \end{aligned}$$

This completes the proof. □

Special score for case-control design II. For case-control design II, we need a similar result.

Result 2 *Consider the model $O^* = (M, W, A, Y) \sim P_0^* \in \mathcal{M}^*$ and let $T^*(P^*)$ denote the tangent space at $P^* \in \mathcal{M}^*$ and assume it satisfies orthogonal decomposition (9). Given a $D^* \in T^*(P^*)$, we have*

$$\begin{aligned} S_m^*(M, W, A, Y) &\equiv I(M = m) q_0(Y \mid m) \{D^*(m, W, A, Y) - E^*(D^* \mid M, Y)\} \\ &\in T^*(P^*). \end{aligned} \tag{10}$$

The same result applies if we replace $q_0(0 \mid m)$ by $q_0(0 \mid m)/J$.

Proof. Firstly, we note that for each m, δ, $\Pi(D^* \mid T_{m,\delta}(P^*)) \in T^*(P^*)$, and by linearity of the space $T_{m,\delta}(P^*)$ (i.e., closure under multiplication by scalar) we have that $q_0(\delta \mid m)\Pi(D^* \mid T^*_{m,\delta}(P^*)) \in T^*(P^*)$. By linearity of $T^*(P^*)$, it follows thus that

$$\begin{aligned}
&\textstyle\sum_\delta q_0(\delta \mid m)\Pi(D^* \mid T^*_{m,\delta}(P^*)) \\
&= \textstyle\sum_\delta q_0(\delta \mid m) I(M = m, Y = \delta)\,(D^*(m, W, A, \delta) - E^*(D^* \mid M, Y)) \\
&= I(M = m) q_0(Y \mid m)\,(D^*(m, W, A, Y) - E^*(D^* \mid M, Y)) \\
&= S^*_m(M, W, A, Y) \\
&\in T^*(P^*).
\end{aligned}$$

This completes the proof. □

Case-control weighted score equals a score, case-control design I. We are now ready to establish our wished results showing that the case-control weighted canonical gradient of the prospective sampling model is an element of the tangent space for the observed data model $\mathcal{M}$.

Theorem 12 *(***Case-control weighted score is a score, Design I***) Consider case-control design I, its independence model $\mathcal{M}$ described by (3), and assume the tangent space $T^*(P^*)$ of $\mathcal{M}^*$ at P^* satisfies the orthogonal decomposition (7).*

If $D^* \in T^*(P^*)$, then

$$D_{q_0}(O) = q_0 D^*(W_1, A_1, 1) + \frac{(1 - q_0)}{J} \sum_j D^*(W_2^j, A_2^j, 0) \in T_I(P^*).$$

Specifically, if we set

$$S^*(W, A, Y) = q_0(Y)\,\{D^*(W, A, Y) - E^*(D^* \mid Y)\} \in T^*(P^*),$$

where $q_0(Y) = I(Y = 1)q_0 + I(Y = 0)(1 - q_0)/J$, then

$$\begin{aligned}
D_{q_0}(O) \;=\; & S^*(W_1, A_1, 1) - E^*(S^*(W, A, Y) \mid Y = 1) \\
& + \sum_j \{S^*(W_2^j, A_2^j, 0) - E^*(S^*(W, A, Y) \mid Y = 0)\}.
\end{aligned}$$

(Here, we use the fact for $J = 1$, $E^*(S^* \mid Y = 1) + E^*(S^* \mid Y = 0) = 0$.)

This establishes the wished corollary stating that the case-control weighted canonical gradient for the prospective sampling model yields the canonical gradient for the case-control sampling model $\mathcal{M}$.

Corollary 1 *Consider case-control design I, its independence model $\mathcal{M}$ described by (3), and assume the tangent space $T^*(P^*)$ of $\mathcal{M}^*$ at P^* satisfies the orthogonal decomposition (7).*

Suppose that $D^*(P^*)$ is the canonical gradient of $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d \Psi^*$, and let $\Psi : \mathcal{M} \to \mathbb{R}^d$ at $P(P^*) \in \mathcal{M}$, satisfy $\Psi(P(P^*)) = \Psi^*(P^*)$.

Assume that the corresponding case-control weighted D_{q_0} (satisfies the regularity conditions such that it) is a gradient for Ψ at $P(P^*)$. Then D_{q_0} is the canonical gradient of Ψ at $P(P^*)$.

Case-control weighted score is a score, Case-Control Design II. We establish the same type result for case-control design II.

Theorem 13 *(***Case-control weighted score is a score, Design II***) Consider case-control design II, its independence model $\mathcal{M}$ described by (4), and assume the tangent space $T^*(P^*)$ of $\mathcal{M}^*$ at P^* satisfies the orthogonal decomposition (9).*

For any $D^* \in L^2(P^*)$, we have

$$\begin{aligned} D_{q_0,\bar{q}_0}(O) &\equiv q_0 D^*(M_1, W_1, A_1, 1) + \bar{q}_0(M_1)\frac{1}{J}\sum_j D^*(M_1, W_2^j, A_2^j, 0) \\ &= \sum_m \frac{q_0}{q_0(1 \mid m)} I(M_1 = m) D^*_{m,q_0}, \end{aligned}$$

where

$$D^*_{m,q_0}(O) \equiv q_0(1 \mid m) D^*(m, W_1, A_1, 1) + \frac{q_0(0 \mid m)}{J} D^*(m, W_2^j, A_2^j, 0).$$

For each m, and $D^* \in T^*(P^*)$, we have

$$I(M_1 = m) D^*_{m,q_0} \in T_{II}(P^*)$$

so that it follows that

$$D_{q_0,\bar{q}_0}(P^*) \in T_{II}(P^*).$$

Let $q_{0J}(\delta \mid m) = q_0(1 \mid m)\delta + (1-\delta)q_0(0 \mid m)/J$. Specifically, if we set

$$S^*_m(M, W, A, Y) = I(M = m) q_{0J}(Y \mid m) \left\{ D^*(m, W, A, Y) - E^*(D^* \mid M, Y) \right\},$$

which is an element of $T^*(P^*)$ by (10) above, then

$$\begin{aligned} I(M_1 = m) D^*_{m,q_0}(O) &= S^*_m(M_1, W_1, A_1, 1) - E^*(S^*_m \mid M, Y = 1) \\ &\quad + \sum_j \{ S^*_m(M_1, W_2^j, A_2^j, 0) - E(S^* \mid M, Y = 0) \} \\ &\in T_{II}(P^*). \end{aligned}$$

Here we use that for any $D^* \in L_0^2(P^*)$,

$$q_0(1 \mid m)E^*(D^* \mid M = m, Y = 1) + q_0(0 \mid m)E(D^* \mid M = m, Y = 0) = 0.$$

This gives us the wished result.

Corollary 2 *(***Case-control weighted canonical gradient is a canonical gradient, Design II***) Consider case-control design II, its independence model* $\mathcal{M}$ *described by (4), and assume the tangent space* $T^*(P^*)$ *of* $\mathcal{M}^*$ *at* P^* *satisfies the orthogonal decomposition (9).*

If $D^*(P^*)$ *is the canonical gradient of* $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ *at* P^**, then*

$$\begin{aligned} D_{q_0,\bar{q}_0} &\equiv \sum_m \frac{q_0}{q_0(1 \mid m)} I(M_1 = m) D^*_{m,q_0} \\ &\in T_{II}(P^*). \end{aligned}$$

Thus, under the conditions for which which $D_{q_0,\bar{q}_0}(P^*)$ *is a gradient of* $\Psi : \mathcal{M} \to \mathbb{R}^d$ *at* $P(P^*) \in \mathcal{M}$*, satisfying* $\Psi(P(P^*)) = \Psi^*(P^*)$ *for specified parameter* $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$*, we also have that* $D_{q_0,\bar{q}_0}(P^*)$ *is the canonical gradient of* Ψ *at* $P(P^*)$.

References

J.A. Anderson. Separate sample logistic discrimination. *Biometrika*, 59:19–35, 1972.

J. Benichou and S. Wacholder. A comparison of three approaches to estimate exposure-specific incidence rates from population-based case-control data. *Statistics in Medicine*, 13:651–661, 1994.

P.J. Bickel, C.A. J. Klaassen, Y. Ritov, and J.A. Wellner. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press, Baltimore, MD, 1993. ISBN 0-8018-4541-6.

N.E. Breslow. Statistics in epidemiology: the case-control study. *J Am Stat Soc*, 91:14–28, 1996.

N.E. Breslow and N.E. Day. *Statistical Methods in Cancer Research: Volume 1 – The analysis of case-control studies*. International Agency for Research on Cancer, Lyon, 1980.

N.E. Breslow, N.E. Day, K.T. Halvorsen, R.L. Prentice, and C. Sabal. Estimation of multiple relative risk functions in matched case-control studies. *Am J Epid*, 108(4):299–307, 1978.

N.E. Breslow, J.M. Robins, and J.A. Wellner. On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli*, 6(3): 447–455, 2000.

D. Collett. *Modeling Binary Data*. Chapman and Hall, London, 1991.

J. Cornfield. A method of estimating comparative rates from clinical data. applications to cancer of the lung, breast, and cervix. *J Nat Cancer Inst*, 11:1269–1275, 1951.

J. Cornfield. A statistical problem arising from retrospective studies. In J. Neyman, editor, *Proceedings of the Third Berkeley Symposium, Volume IV*, pages 133–148. University of California Press, 1956.

S.R. Cosslett. Maximum likelihood estimator for choice-based samples. *Econometrica*, 49(5):1289–1316, 1981.

S. Greenland. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *Am J Epidemiol*, 160(4):301–305, 2004.

S. Greenland. Multivariate estimation of exposure-specific incidence from case-control studies. *J Chron Dis*, 34:445–453, 1981.

T.R. Holford, C. White, and J.L. Kelsey. Multivariate analysis for matched case-control studies. *Am J Epid*, 107(3):245–255, 1978.

D.W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. John Wiley and Sons, New York, 2nd edition, 2000.

N.P. Jewell. *Statistics for Epidemiology*. Chapman and Hall/CRC, Boca Raton, 2004.

C.F. Manski and S.R. Lerman. The estimation of choice probabilities from choice based samples. *Econometrica*, 45(8):1977–1988, 1977.

C.F. Manski and D. McFadden. Alternative estimators and sample designs for discrete choice analysis. In C.F. Manski and D. McFadden, editors, *Structural Analysis of Discrete Data with Econometric Applications*. The MIT Press, Cambridge, MA, 1981.

R. Mansson, M.M. Joffe, W. Sun, and S. Hennessy. On the estimation and use of propensity scores in case-control and cohort studies. *American Journal of Epidemiology*, 00:1–8, 2007.

K.L. Moore and M.J. van der Laan. Covariate adjustment in randomized trials with binary outcomes. Technical report 215, Division of Biostatistics, University of California, Berkeley, April 2007.

A.P. Morise, G.A. Diamon, R. Detrano, M. Bobbio, and Erdogan Gunel. The effect of disease-prevalence adjustments on the accuracy of a logistic prediction model. *Med Decis Making*, 16:133–142, 1996.

S. Newman. Causal analysis of case-control data. *Epid Persp Innov*, 3:2, 2006. URL `http://www.epi-perspectives.com/content/3/1/2`.

R.L. Prentice and N.E. Breslow. Retrospective studies and failure time models. *Biometrika*, 65(1):153–158, 1978.

R.L. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66:403–411, 1979.

J.M. Robins. [choice as an alternative to control in observational studies]: Comment. *Statistical Science*, 14(3):281–293, 1999.

J.M. Robins, A. Rotnitzky, and L.P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, September 1994.

K. Rothman and S. Greenland. *Modern Epidemiology*. Lippincott, Williams and Wilkins, Philadelphia, PA, 2nd edition, 1998.

D.B. Rubin. *Matched Sampling for Causal Effects*. Cambridge University Press, Cambridge, MA, 2006.

J.J. Schlesselman. *Case-Control Studies: Design, Conduct, Analysis*. Oxford University Press, Oxford, 1982.

M.J. van der Laan. Causal effect models for intention to treat and realistic individualized treatment rules. Technical report 203, Division of Biostatistics, University of California, Berkeley, 2006.

M.J. van der Laan and J.M. Robins. Unified methods for censored longitudinal data and causality. Springer, New York, 2002.

M.J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

S. Wachholder. The case-control study as data missing by design: Estimating risk differences. *Epidemiology*, 7(2):144–150, 1996.

8.2 *Simple Optimal Weighting of Cases and Controls in Case-Control Studies*

The following article appears as it was published in the *International Journal of Biostatistics* in 2008, http://www.bepress.com/ijb/vol4/iss1/19/.

It was originally published on the University of California, Berkeley Division of Biostatistics Working Paper Series website as *A Guide to Causal Parameters in Case-Control Designs: Targeted Maximum Likelihood Estimation* in 2008,
http://www.bepress.com/ucbbiostat/paper235/.

Simple Optimal Weighting of Cases and Controls in Case-Control Studies

Sherri Rose and Mark J. van der Laan
Division of Biostatistics, University of California, Berkeley
sherri@berkeley.edu, laan@berkeley.edu

Abstract

Researchers of uncommon diseases are often interested in assessing potential risk factors. Given the low incidence of disease, these studies are frequently case-control in design, as this allows for a sufficient number of cases to be obtained without extensive sampling and can increase efficiency. However, these case-control samples are then biased since the proportion of cases in the sample is not the same as the population of interest. Methods for analyzing case-control studies have focused on utilizing logistic regression models that provide conditional and not causal estimates of the odds ratio. This article will demonstrate the use of the prevalence probability and case-control weighted targeted maximum likelihood estimation (MLE), as described by van der Laan (2008), in order to obtain causal estimates of the parameters of interest (risk difference, relative risk, and odds ratio). It is meant to be used as a guide for researchers, with step-by-step directions to implement this methodology. We will also present simulation studies that show the improved efficiency of the case-control weighted targeted MLE compared to other techniques.

1 Introduction

Case-control study designs are frequently used in public health and medical research to assess potential risk factors for disease. These study designs are particularly attractive to investigators researching rare diseases (i.e. probability of having the disease ≈ 0), as they are able to sample known cases of disease, versus following a large number of subjects and waiting for disease onset. Case-control studies can also yield increases in efficiency. However, case-control sampling is a biased sampling method; bias occurs due to the disproportionate number of cases in the sample versus the population of interest. Researchers commonly employ the use of a logistic regression model, treating the sample as a prospective sample, and estimate the *conditional* odds ratio of having disease given the exposure of interest and measured covariates. If one would like to estimate *marginal causal* effects, which correspond with the traditional parameters of interest in randomized trials, there is now a nonparametric double robust locally efficient procedure available. In van der Laan (2008), methodology for this marginal causal effect estimation theory in case-control designs is illustrated in detail. These techniques rely on knowledge of the true prevalence probability $P_0^*(Y = 1) \equiv q_0$ to eliminate the bias of the case-control sampling design. This methodology can be used in conjunction with other available procedures that handle censoring, missingness, measurement error, and other issues that are persistent in prospective and retrospective studies.

When possible, the population under study should be clearly defined. As such, the prevalence and incidence probabilities are then truly basic information about a population of interest. Given the availability of city, state, and national databases for many diseases, including many cancers, knowledge of the prevalence and incidence probabilities is now increasingly realistic. The literature, going back to the 1950's, supports this (see Cornfield (1951) and Cornfield (1956)). Nested case-control studies can also take advantage of the prevalence or incidence probability available in the full cohort study. The appropriateness of the use of the prevalence versus the incidence probability will depend on the nature of the case-control study design. The use of the these probabilities to eliminate the bias of case-control sampling design has previously been discussed as update to a logistic regression model with the intercept $\log q_0/(1-q_0)$ (Anderson, 1972; Prentice and Breslow, 1978; Greenland, 1981; Morise et al., 1996; Wacholder, 1996; Greenland, 2004). When the appropriate probability, or an estimate of the probability, is available, our procedure (van der Laan, 2008) can be implemented. In situations where data on the population of interest may be sparse, the use of a range for the probability is

still beneficial.

An existing method for causal inference in case-control study designs, discussed by Robins (1999) and Mansson et al. (2007), involves the use of the exposure mechanism (also known as the propensity score or treatment mechanism in other literature) among control subjects as a weight to update a logistic regression of disease status on exposure. This inverse probability of treatment weighted (IPTW) marginal structural model does not require the knowledge of prevalence probability, only that the prevalence probability is close to zero. We will discuss this and other existing methods for analysis of case-control studies while stressing our new case-control weighting method that utilizes the prevalence probability.

The procedure, case-control weighted targeted maximum likelihood estimation, "targets" the parameter of interest rather than the distribution of interest. We use extra information, the estimate of the conditional distribution of the exposure given covariates among cases and controls (which we refer to as the exposure mechanism), to update an initial estimate of $P_0^*(Y \mid A, W)$. The procedure is double robust and locally efficient: it performs well as long as $P_0^*(Y \mid A, W)$ or $P_0^*(A \mid W)$ is correctly specified, is consistent if either of these models are correctly specified, and efficient if both are correctly specified. Our case-control weighting scheme successfully maps estimation methods designed for prospective sampling into methods for case-control sampling. It also produces efficient estimators when its prospective sample counterpart is efficient. For theoretical development of this new methodology, we will refer to van der Laan (2008). This article discusses case-control weighted targeted maximum likelihood for cumulative study designs with the prevalence probability and will focus on applications of the case-control weighting scheme in unmatched (independent) studies. For an extension of the methodology to matched case-control studies, see van der Laan (2008) and Rose and van der Laan (2008). Theory for incidence-density sampling with the incidence probability is also presented as an appendix in van der Laan (2008).

1.1 Case-Control Estimation

For ease of reference throughout the remainder of this article, we will present basic notation for understanding of the case-control estimation problem here. Let us define $O^* = (W, A, Y) \sim P_0^*$ as the experimental unit and corresponding distribution P_0^* of interest, which consists of baseline covariates W, an exposure variable A, and a binary outcome Y that defines case or control status. (For prospective studies, the exposure variable A would be referred to as the "treatment" variable.) P_0^* therefore represents the population from which all

cases and controls will be sampled. One might be interested in several marginal causal effect parameters, including the causal risk difference, relative risk, and odds ratio. For causal effect parameter $\psi_0^* = \Psi^*(P_0^*) \in \mathbb{R}^d$ of $P_0^* \in \mathcal{M}^*$ and binary exposure $A \in \{0,1\}$, these parameters are defined as:

$$\begin{aligned}\psi_{0,RD}^* &\equiv E_0^*\{E_0^*(Y \mid A=1,W) - E_0^*(Y \mid A=0,W)\} \\ &= E_0^*(Y_1) - E_0^*(Y_0) \\ &= P_0^*(Y_1=1) - P_0^*(Y_0=1), \end{aligned} \tag{1}$$

$$\psi_{0,RR}^* = \frac{E_0^*E_0^*(Y \mid A=1,W)}{E_0^*E_0^*(Y \mid A=0,W)} = \frac{E_0^*(Y_1)}{E_0^*(Y_0)} = \frac{P_0^*(Y_1=1)}{P_0^*(Y_0=1)}, \tag{2}$$

and,

$$\psi_{0,OR}^* = \frac{P_0^*(Y_1=1)P_0^*(Y_0=0)}{P_0^*(Y_1=0)P_0^*(Y_0=1)}, \tag{3}$$

respectively. The causal versions of these definitions require the specification of the counterfactual outcomes Y_0 and Y_1 for binary A and $(W, A, Y = Y_A)$ as a time-ordered missing data structure on (W, Y_0, Y_1), the full data structure. In addition, one must make the randomization assumption: $\{A \perp Y_0, Y_1 \mid W\}$. On the other hand, these parameters are always well defined parameters of the distribution of the data, and they can thereby be viewed as W-adjusted variable importance parameters without the need to make these assumptions. See van der Laan (2006) for this framework.

In van der Laan (2008), independent case-control sampling is described as first sampling (W_1, A_1) from the conditional distribution of (W, A), given $Y = 1$ for a case and then sampling J controls (W_0^j, A_0^j) from (W, A), given $Y = 0, j = 1, \ldots, J$. The observed data structure in independent case-control sampling is then defined by:

$$O = ((W_1, A_1), (W_0^j, A_0^j : j = 1, \ldots, J)) \sim P_0, \text{ with}$$

$$(W_1, A_1) \sim (W, A \mid Y = 1)$$

$$(W_0^j, A_0^j) \sim (W, A \mid Y = 0)$$

where the cluster containing one case and J controls is considered the experimental unit, and the marginal distribution of this cluster is specified by P_0^*. Therefore, a case-control data set consists of n independent and identically distributed observations $O_1, \ldots, O_n$ with sampling distribution P_0 as described

above. The model $\mathcal{M}^*$, where q_0 may or may not be known, implies models for the marginal distribution of cases (W_1, A_1) and controls $(W_2^j, A_2^j), j = 1, \ldots, J$.

This coupling formulation was useful when proving results for the case-control weighting methodology, and the tools provided in van der Laan (2008) show that the following is also true. If independent case-control sampling is described as sampling nC cases from the conditional distribution of (W, A), given $Y = 1$, and sampling nCo controls from (W, A), given $Y = 0$, the value of J used to weight each control is then nCo/nC. This simple ratio $J = nCo/nC$ can be used effectively in practice.

2 Existing Methodology

As previously discussed, conditional estimation of the odds ratio of being diseased given the exposure of interest and baseline covariates is the prevalent method of analysis in case-control study designs. Key publications in the area of logistic regression for independent case-control study designs are Anderson (1972), Prentice and Pyke (1979), Breslow and Day (1980), and Breslow (1996). Greenland (1981) and Holland and Rubin (1988) discuss another model-based method: the use of log-linear models to estimate the marginal odds ratio. There are also several references for standardization in case-control studies, which estimates marginal effects with population or person-time averaging, including Rothman and Greenland (1998) and Greenland (2004). Benichou and Wacholder (1994) also present multivariate methods for population-based case-control studies. In this section, we will discuss the use of an intercept adjusted logistic regression as it can be incorporated into our case-control weighting framework. We will also discuss an IPTW marginal structural model for the estimation of causal effects as it is a related methodology, making use of the exposure mechanism. While these methods are discussed in current literature, they are infrequently implemented in current public health and medical research compared to the use of logistic regression for conditional effects.

2.1 Intercept Adjusted Logistic Regression

A thorough literature search yielded several publications suggesting the use of $\log q_0/(1 - q_0)$ as an update to the intercept of a logistic regression. (See Anderson (1972), Prentice and Breslow (1978), Greenland (1981), Morise et al. (1996), Wacholder (1996), and Greenland (2004), among others.) However, its use in practice remains limited. The adjustment is sometimes presented as a

ratio of sampling fractions:

$$\log\Big(\frac{P(\text{sampled} \mid Y=1)}{P(\text{sampled} \mid Y=0)}\Big),$$

which reduces to $\log q_0/(1-q_0)$.

Adding the intercept $\log q_0/(1-q_0)$, denoted as $\log c_0$, yields the true logistic regression function $P_0^*(Y=1 \mid A,W)$ (Anderson, 1972; Prentice and Pyke, 1979). An intercept adjusted logistic regression can be used within the case-control weighting framework as an initial estimate of $P_0^*(Y \mid A,W)$. This will be discussed further in Section 3.2.1 and Section 4. The true logistic regression function can also be mapped to causal effect parameters by averaging over the case-control weighted distribution of W, which will also be discussed in Section 3.2.1.

2.2 IPTW

Robins (1999) and Mansson et al. (2007) discuss, under a rare disease assumption, the use of an approximately correct IPTW method in a marginal structural logistic regression model for case-control study designs. This procedure uses the estimated propensity score (exposure mechanism) among control subjects to update a logistic regression of Y on A. However, this IPTW estimator targets a nonparametrically non-identifiable parameter, which indicates strong sensitivity towards model misspecification for the exposure mechanism. See van der Laan (2008) for formal discussion of this result. Additionally, the causal effect estimates of the risk difference and relative risk cannot be obtained using this method. We also refer to Newman (2006) for a related IPTW-type approach for fitting marginal structural models based on case-control data. This method builds on the standardization approach in order to weight exposed and unexposed controls using a regression of A on W. We will include the IPTW method of Robins (1999) and Mansson et al. (2007) in our simulations.

3 Case-Control Weighted Targeted Maximum Likelihood Estimation

In this section, we provide the end user with a practical overview of the case-control weighting scheme for targeted maximum likelihood estimation in case-control study designs. For the formal statistical theory behind this technique,

see van der Laan (2008). We discuss the implementation of case-control weighting for targeted maximum likelihood estimation both broadly and step-wise so that this article may be used as a guide to researchers wishing to employ these methods in their work.

3.1 Summary

Case-control weighted targeted maximum likelihood estimation for case-control study designs differs from other approaches to causal parameter estimation in case-control study design as it incorporates estimates of $P_0^*(Y \mid A, W)$, $P_0^*(A \mid W)$, and knowledge of q_0. Intercept adjusted logistic regression mapped to causal parameters discussed in the previous section relies on knowledge of only $P_0^*(Y \mid A, W)$ and q_0; the IPTW procedure of Robins (1999) and Mansson et al. (2007) relies on $P_0^*(A \mid W)$. The case-control weighted targeted maximum likelihood estimation procedure provides a nonparametric double robust locally efficient estimator: it performs well as long as $P_0^*(Y \mid A, W)$ or $P_0^*(A \mid W)$ is correctly specified, is consistent if either of these models are correctly specified, and efficient if both are correctly specified. It uses extra information, the estimate of the conditional distribution of the exposure given covariates among cases and controls, to update an initial estimate of $P_0^*(Y \mid A, W)$. One can use data-adaptive model-selection for estimation of $P_0^*(Y \mid A, W)$ and $P_0^*(A \mid W)$ within our procedure. (This will be discussed further in Section 5.) The procedure follows the basic steps enumerated below, which we then illustrate in more detail.

1. Assign weights q_0 to the cases and $(1 - q_0)\frac{1}{J}$ to the corresponding J controls.

2. Estimate the conditional probability of Y given A and W using assigned weights. The estimate of $P_0^*(Y \mid A, W) \equiv Q_0^*(A, W)$ is $\hat{Q}^*(A, W)$.

3. Estimate the conditional distribution of the exposure given covariates using assigned weights. The estimate of $P_0^*(A \mid W) \equiv g_0^*(A \mid W)$ is $\hat{g}^*(A \mid W)$.

4. Calculate the "clever covariate" for each subject based on $g_0^*(A \mid W)$. The covariate is estimated by $h(A, W)$.

5. Update the initial fit $\hat{Q}^*(A, W)$ from step 2 using the covariate $h(A, W)$. This is achieved by holding the coefficients of $\hat{Q}^*(A, W)$ fixed while estimating a new coefficient ϵ for $h(A, W)$ using weighted maximum likelihood estimation. The updated regression is given by $\hat{Q}_1^*(A, W)$

6. Use the assigned weights and $\hat{Q}_1^*(A, W)$ to estimate causal parameters of interest seen in formulas (1), (2) and (3). This is done by averaging over the case-control weighted distribution of W.

7. Calculate standard errors, and then, subsequently, p-values and confidence intervals, using the influence curve.

3.2 Implementation

The implementation of case-control weighted targeted maximum likelihood can be achieved using existing tools available in current software packages. Here we illustrate the steps described in Section 3.1.

3.2.1 Estimating $Q_0^*(A, W)$

After assigning weights q_0 and $(1-q_0)\frac{1}{J}$ to cases and controls, respectively, the first step in case-control weighted targeted maximum likelihood estimation for case-control designs is obtaining an estimate for $P_0^*(Y \mid A, W) \equiv Q_0^*(A, W)$. We offer two approaches for fitting this initial regression, the previously discussed intercept adjusted logistic regression, and a case-control weighted logistic regression. A comparison of these two approaches will be discussed in Section 4.

Intercept Adjusted Logistic Regression for $Q_0^*(A, W)$. Updating a logistic regression with $\log c_0$ is discussed in Section 2.1.

Case-Control Weighted Logistic Regression for $Q_0^*(A, W)$. Using the assigned weights, one simply performs maximum likelihood estimation for prospective sampling ignoring the case-control sampling design. If one considers a nonparametric model for the marginal distribution of the covariates and a model $\{Q_\theta^* : \theta\}$ for $Q_0^*(A, W)$, the case-control weighted maximum likelihood estimator for $Q_0^*(A, W)$ is then given by:

$$\hat{\theta} = \arg\max_\theta \sum_{i=1}^{n} q_0 \log \hat{Q}_\theta^*(A_{1i}, W_{1i}) + (1 - q_0)\frac{1}{J}\sum_{j=1}^{J} \log(1 - \hat{Q}_\theta^*(A_{2i}^j, W_{2i}^j)).$$

Implementing case-control weighted maximum likelihood estimation, which is simply a weighted logistic regression, is quite straightforward, and can be done in many existing statistical software programs, including `SAS`, `STATA`, and `R`.

Outside of the case-control weighted targeted maximum likelihood estimation framework, case-control weighted logistic regression mapped to causal

inference parameters produce efficient estimators. This mapping is accomplished by evaluating $\hat{Q}^*(A,W)$ at $A=1$ and $A=0$, applying the appropriate weights to estimate $P_0^*(Y_1=1)$ and $P_0^*(Y_0=1)$, and then computing the desired causal parameters of interest defined in formulas (1), (2), and (3). Estimating causal parameters will be discussed in more detail in Section 3.2.5. Case-control weighted logistic regression therefore provides researchers an immediate one-step intuitive procedure to estimate causal inference parameters in case-control study designs.

3.2.2 Estimating $g_0^*(A \mid W)$

The case-control targeted maximum likelihood estimation procedure uses the estimate of $Q_0^*(A,W)$ obtained above in conjunction with an estimate of $g_0^*(A \mid W)$. If one further considers a model $\{g_\eta^* : \eta\}$ for $g_0^*(A \mid W)$, the case-control weighted maximum likelihood estimator for $g_0^*(A \mid W)$ is given by:

$$\hat{\eta} = \arg\max_{\eta} \sum_{i=1}^{n} q_0 \log \hat{g}_\eta^*(A_{1i} \mid W_{1i}) + (1-q_0)\frac{1}{J}\sum_{j=1}^{J} \log \hat{g}_\eta^*(A_{2i}^j \mid W_{2i}^j),$$

For improved performance of the targeted maximum likelihood estimator in a practical environment, estimated probabilites that are smaller than 0.01 can be set to 0.01 (Bembom et al., 2007).

3.2.3 Calculating $h(A,W)$

After estimating $Q_0^*(A,W)$ and $g_0^*(A \mid W)$, the next step requires calculation of a "clever covariate" for each subject. This covariate, which is calculated as if one has a prospective sample, takes the form:

$$h(A,W) \equiv \left(\frac{I(A=1)}{\hat{g}^*(A=1 \mid W)} - \frac{I(A=0)}{\hat{g}^*(A=0 \mid W)}\right)$$

for the risk difference. It is easy to see that for $A=1$ the second term disappears, and for for $A=0$ the first term disappears. Two covariates:

$$h_0(A,W) \equiv \left(-\frac{I(A=0)}{\hat{g}^*(A=0 \mid W)}\right) \text{ and } h_1(A,W) \equiv \left(\frac{I(A=1)}{\hat{g}^*(A=1 \mid W)}\right)$$

are used for estimation of other parameters, such as the relative risk and odds ratio. For a more detailed discussion of the "clever covariate," see van der Laan and Rubin (2006) and Moore and van der Laan (2007).

3.2.4 Updating $\hat{Q}^*(A, W)$

Updating $\hat{Q}^*(A, W)$ involves performing an additional weighted regression with $h(A, W)$ as a supplementary covariate. All other coefficients in the initial regression $\hat{Q}^*(A, W)$ are held fixed, and an intercept is suppressed in order to estimate the coefficient in front of $h(A, W)$, denoted ϵ. The case-control weighted targeted maximum likelihood estimation procedure is then able to incorporate information from $\hat{g}^*(A \mid W)$, through $h(A, W)$, into an updated regression. It does this by extracting $\hat{\epsilon}^1$, the case-control weighted maximum likelihood estimator of ϵ, from the fit defined above, and updating the regression estimate $\hat{Q}^*(A, W)$. This updated regression is then given by $\hat{Q}_1^*(A, W)$:

$$\hat{Q}_1^*(A, W) = \hat{Q}^*(A, W) + \hat{\epsilon}^1 h(A, W).$$

The updating procedure is iterated until convergence, although in many examples convergence is achieved in one step.

3.2.5 Estimating Causal Parameters

The risk difference, relative risk, and odds ratio, were previously defined generally in formulas (1), (2), and (3). The estimate $\hat{Q}_1^*(A, W)$ obtained in the previous step can be easily mapped into causal parameters of interest in the case-control weighting scheme for targeted maximum likelihood estimation by averaging over the case-control weighted distribution of W. This is accomplished by evaluating $\hat{Q}_1^*(A, W)$ at $A = 1$ and $A = 0$ and applying weights q_0 for cases and $(1 - q_0)\frac{1}{J}$ to the corresponding J controls to form case-control weighted estimates of $E_0^*(Y_1) = P_0^*(Y_1 = 1)$ and $E_0^*(Y_0) = P_0^*(Y_0 = 1)$. The risk difference, relative risk, and odds ratio can then be simply calculated from these estimates. For example, the relative risk $E_0^*(Y_1)/E_0^*(Y_0)$ is estimated by:

$$\hat{\psi}_{RR} = \frac{\frac{1}{n}\sum_{i=1}^n q_0 \hat{Q}_{1,q_0}^*(1, W_{1i}) + (1 - q_0)\frac{1}{J}\sum_j \hat{Q}_{1,q_0}^*(1, W_{2i}^j)}{\frac{1}{n}\sum_{i=1}^n q_0 \hat{Q}_{1,q_0}^*(0, W_{1i}) + (1 - q_0)\frac{1}{J}\sum_j \hat{Q}_{1,q_0}^*(0, W_{2i}^j)}.$$

3.2.6 Calculating Standard Errors

The calculation of standard errors for case-control weighted targeted maximum likelihood involves the use of case-control weighted influence curves for the risk difference, relative risk, and odds ratio. This methodology is discussed in detail in van der Laan (2008), and a complete technical understanding of infuence curve derivation is not necessary to implement the case-control targeted maximum likelihood estimation procedure. We also refer to van der

Laan and Robins (2002) for careful discussions of gradients and influence curve theory.

For example, the unweighted influence curve for the risk difference of a prospective study $\psi^*_{0,RD} = P^*_0(Y_1 = 1) - P^*_0(Y_0 = 1)$ is estimated by:

$$\begin{aligned}\hat{D}_{RD}(\psi^*, g^*, Q^*)(O) &= \frac{I(A=1)}{\hat{g}^*(1 \mid W)}(Y - \hat{Q}^*(1, W)) - \frac{I(A=0)}{\hat{g}^*(0 \mid W)}(Y - \hat{Q}^*(0, W)) \\ &\quad + \hat{Q}^*(1, W) - \hat{Q}^*(0, W) - \hat{\psi}.\end{aligned}$$

The case-control weighted influence curve for the risk difference $\psi^*_{0,RD} = P^*_0(Y_1 = 1) - P^*_0(Y_0 = 1)$ is then estimated by:

$$\begin{aligned}\hat{D}_{RD,q_0}(\psi^*, g^*, Q^*)(O) &= q_0 \hat{D}^*(g^*, Q^*)(A_1, W_1, 1) \\ &\quad + (1 - q_0)\frac{1}{J}\sum_{j=1}^{J} \hat{D}^*(g^*, Q^*)(A_2^j, W_2^j, 0) - \hat{\psi}.\end{aligned}$$

Note that the case-control weighted influence curve is merely the influence curve for prospective targeted maximum likelihood with case-control weighting. See van der Laan and Rubin (2006) and Moore and van der Laan (2007) for prospective sampling targeted maximum likelihood methodology.

An estimate of the asymptotic variance of $\sqrt{n}(\hat{\psi} - \psi^*_0)$ using the estimate of the efficient influence curve $D_{q_0}(\psi^*, g^*, Q^*)(O)$ is given by:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} D^2_{q_0}(\psi^*, g^*, Q^*)(O).$$

Given the influence curve for the causal parameter estimate $\hat{\psi}$, a 95% Wald-type confidence interval can be constructed as: $\hat{\psi} \pm z_{0.975}\frac{\hat{\sigma}}{\sqrt{n}}$. Likewise, the p-value of $\hat{\psi}$ can be calculated as $2[1 - \Phi(|\frac{\hat{\psi}}{\hat{\sigma}/\sqrt{n}}|)]$.

4 Intercept Adjusted MLE and Case-Control Weighted MLE

Intercept adjusted maximum likelihood estimation and case-control weighted maximum likelihood estimation were previously discussed as options for the initial fit $\hat{Q}^*(A, W)$. Several issues became apparent when using intercept adjusted maximum likelihood estimation for $\hat{Q}^*(A, W)$ in our case-control weighted targeted maximum likelihood framework. In multiple simulation settings we found that when $\hat{Q}^*(A, W)$ was misspecified using an intercept

adjusted fit, the predicted probabilities were substantially biased compared to the misspecified case-control weighted maximum likelihood probabilties. This additional bias can be understood intuitively since the update to the logistic regression $\log c_0$ is static regardless of the model used, and the parameters of the model (excluding the intercept) are not adjusted by this update. For correctly specified $\hat{Q}^*(A, W)$ this is not an issue, but when $\hat{Q}^*(A, W)$ is misspecified, it leads to substantial bias. Conversely, the case-control weighted logistic regression estimate incorporates the case-control weights each time it fits an estimate. Thus, for misspecified $\hat{Q}^*(A, W)$, case-control weighted predicted probabilities will likely be closer to the truth than intercept adjusted estimates. See Figure 1 for an illustration.

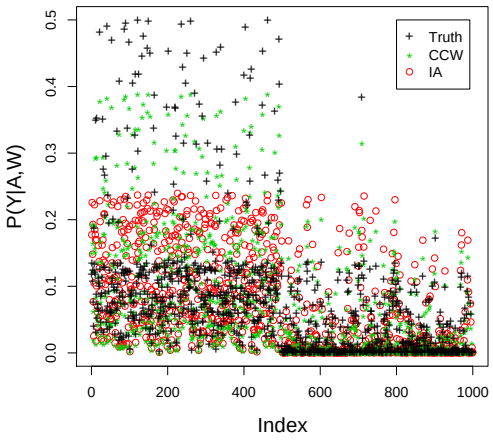


Figure 1: **Predicted Probabilities for Misspecified $\hat{Q}^*(A, W)$.**

The case-control targeted intercept adjusted maximum likelihood improved, with regard to bias, on its non-targeted counterpart for misspecified $\hat{Q}^*(A, W)$. However, the additional bias for misspecified $\hat{Q}^*(A, W)$ and intercept adjusted logistic regression led to much slower convergence to the true values of the risk difference, relative risk, and odds ratio within the case-control targeted maximum likelihood framework. Case-control weighted targeted maximum likelihood with misspecified $\hat{Q}^*(A, W)$ fit with case-control weighted logistic regression became consistent for reasonable sample sizes. Coverage probabilities for case-control weighted targeted intercept adjusted maximum likelihood estimation for misspecified $\hat{Q}^*(A, W)$ also diverged substantially from 95% (as low as 65%) for reasonable sample sizes due to the bias of the estimators. We should note that when $\hat{Q}^*(A, W)$ is correctly specified, the intercept adjusted

methods performed as well as the case-control weighted methods. However, the correct specification of $\hat{Q}^*(A, W)$ is unlikely in practice. Given these findings, we present in our simulations the use of case-control weighted targeted maximum likelihood estimation using case-control weighted logistic regression for the initial fit.

5 Simulation Studies

5.1 Simulation 1

Our first simulation study was designed to illustrate the advantages of the case-control weighting scheme for targeted maximum likelihood estimation in case-control designs. It was based on a population of $N = 120,000$ individuals, where we simulated a 1-dimensional covariate W, a binary exposure A, and indicator Y, which was 1 for cases and 0 for controls. These variables were generated according to the following rules:

$$W \sim U(0,1)$$

$$g_0^*(A \mid W) = \frac{1}{1+\exp(-(W^2-4W+1))}$$

$$Q_0^*(A,W) = \frac{1}{1+\exp(-(1.2A-\sin(W^2)+A\sin(W^2)+5A\log(W)+5\log(W)-1))}.$$

The resulting population had a prevalence probability $q_0 = 0.035$, and exactly $4,165$ cases. We sampled the population using a varying number of cases and controls, and for each sample size we ran 1000 simulations. The true values of the risk difference, relative risk, and odds ratio were given by $RD = 0.043$, $RR = 2.483$, and $OR = 2.598$, with $P(Y_1 = 1) = 0.072$ and $P(Y_0 = 1) = 0.029$. These causal effect parameters were estimated using methods discussed in this paper:

1. **IPTW**: IPTW method for marginal structural models (Robins, 1999; Mansson et al., 2007) that uses the estimated exposure mechanism among the controls to update a logistic regression of Y on A discussed in Section 2.2.

2. **Case-Control Weighted MLE (CCW-MLE)**: Case-control weighted logistic regression, discussed in Section 3.2.1, mapped to causal effect estimators by averaging over the case-control weighted distribution of W.

3. **Case-Control Weighted Targeted MLE (CCW-TMLE)**: Case-control weighted targeted maximum likelihood procedure for case-control designs with case-control weighted $\hat{Q}^*(A,W)$ discussed in Section 3.

The initial fit for each method requiring an estimate of $Q_0^*(A,W)$ was defined by:

$$\hat{Q}^*(A,W) = \frac{1}{1+\exp(-(\hat{\alpha_0}+\hat{\alpha_1}A+\hat{\alpha_2}\log(W)+\hat{\alpha_3}\sin(W^2)+\hat{\alpha_4}A\log(W)+\hat{\alpha_5}A\sin(W^2)))},$$

which was the correctly specified fit. $Q_0^*(A,W)$ was also estimated in a second simulation with:

$$\hat{Q}^*(A,W) = \frac{1}{1+\exp(-(\hat{\alpha_0}+\hat{\alpha_1}A+\hat{\alpha_2}W))},$$

a misspecified fit. For methods requiring a fit for exposure mechanism, the correct fit was defined by:

$$\hat{g}^*(A \mid W) = \frac{1}{1+\exp(\hat{\eta_0}+\hat{\eta_1}W^2+\hat{\eta_2}W)}.$$

The misspecified version of the exposure mechanism was given by:

$$\hat{g}^*(A \mid W) = \frac{1}{1+\exp(\hat{\eta_0}+\hat{\eta_1}W)}.$$

In our simulation study, we realistically generated A dependent on W. This led to some substantial increases in efficiency in the targeted estimator when $\hat{Q}^*(A,W)$ was misspecified and sample size was larger, as they also adjust for $\hat{g}^*(A \mid W)$. This emphasizes the double robustness of the targeted estimators, and suggests that one should always adjust for $\hat{g}^*(A \mid W)$ in practice. When $\hat{Q}^*(A,W)$ was correctly specified, the relative efficiency of the targeted estimator (CCW-TMLE) was similar to its non-targeted counterpart (CCW-MLE), demonstrating that the use of q_0 and $\hat{Q}^*(A,W)$ alone can produce efficient estimators. This was further highlighted in the results for the odds ratio and the IPTW estimators, which do not utilize q_0, as they had the poorest overall efficiency. Mean squared errors and relative efficiencies for the causal odds ratio are provided in Table 1. The results for the relative risk and risk difference are combined in Table 2. The least efficient estimator as sample size increased for these parameters was the case-control weighted logistic regression when $Q_0^*(A,W)$ was realistically misspecified.

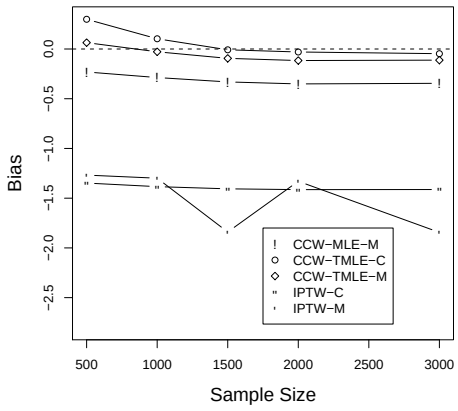
When examining the bias of the estimators for the odds ratio, it is clear that the IPTW estimators had the highest level of bias across all sample sizes,

Table 1: **Simulation 1 – Odds Ratio** – MSE is Mean Squared Error for IPTW Misspecified Estimate, RE is Relative Efficiency of Other Estimators Compared to IPTW Misspecified Estimate MSE, nC is Number of Cases, nCo is Number of Controls, n is Number of Total Observations, M is for Misspecified $\hat{Q}^*(A,W)$ or $\hat{g}^*(A \mid W)$ Fit, C is for Correctly Specified $\hat{Q}^*(A,W)$ or $\hat{g}^*(A \mid W)$. (When two letters are noted in the "Fit" column, the first letter refers to $\hat{Q}^*(A,W)$ and the second to $\hat{g}^*(A \mid W)$.)

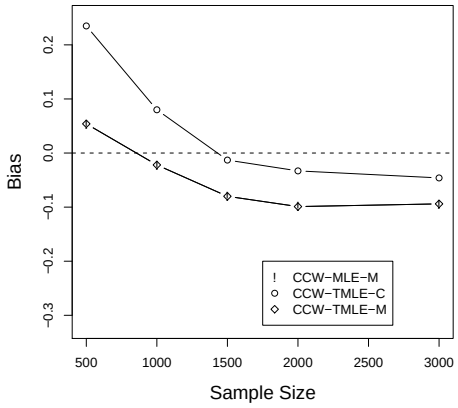
		n=500	n=1000	n=1500	n=2000	n=3000
		nC=250	nC=500	nC=500	nC=1000	nC=1000
Odds Ratio	Fit	nCo=250	nCo=500	nCo=1000	nCo=1000	nCo=2000
IPTW MSE	M	1.76	1.75	3.39	1.80	3.40
IPTW RE	C	0.91	0.89	1.69	0.89	1.69
	C/C	1.27	3.62	14.58	8.40	32.03
CCW-TMLE RE	C/M	1.26	3.62	14.57	8.40	31.97
	M/C	1.96	4.63	16.68	9.52	31.91
CCW-MLE RE	C	1.27	3.65	14.64	8.44	32.12
	M	3.07	5.72	14.54	7.83	18.93

Table 2: **Simulation 1 – Relative Risk and Risk Difference** – MSE is Mean Squared Error for CCW-MLE Misspecified Estimate, RE is Relative Efficiency of Other Estimators Compared to CCW-MLE Misspecified Estimate MSE, nC is Number of Cases, nCo is Number of Controls, n is Number of Total Observations, M is for Misspecified $\hat{Q}^*(A,W)$ or $\hat{g}^*(A \mid W)$ Fit, C is for Correctly Specified $\hat{Q}^*(A,W)$ or $\hat{g}^*(A \mid W)$. (When two letters are noted in the "Fit" column, the first letter refers to $\hat{Q}^*(A,W)$ and the second to $\hat{g}^*(A \mid W)$.)

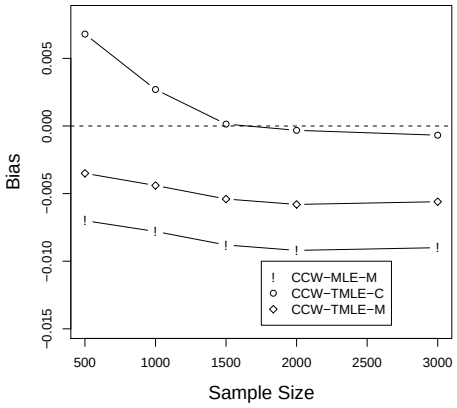
		n=500	n=1000	n=1500	n=2000	n=3000
		nC=250	nC=500	nC=500	nC=1000	nC=1000
Relative Risk	Fit	nCo=250	nCo=500	nCo=1000	nCo=1000	nCo=2000
CCW-MLE MSE	M	0.46	0.25	0.19	0.19	0.15
CCW-MLE RE	C	0.48	0.69	1.06	1.12	1.73
	C/C	0.47	0.68	1.05	1.12	1.73
CCW-TMLE RE	C/M	0.47	0.68	1.05	1.12	1.73
	M/C	0.65	0.82	1.15	1.22	1.69
Risk Difference						
CCW-MLE MSE	M	3.2E-04	1.8E-04	1.4E-04	1.4E-04	1.1E-04
CCW-MLE RE	C	0.45	0.67	1.10	1.15	1.89
	C/C	0.45	0.67	1.10	1.15	1.89
CCW-TMLE RE	C/M	0.45	0.67	1.10	1.15	1.89
	M/C	0.98	1.12	1.34	1.36	1.63



(a) Odds Ratio



(b) Relative Risk



(c) Risk Difference

Figure 2: **Simulation 1 – Bias Results.** (Bias results for the case-control weighted targeted maximum likelihood with misspecified $\hat{g}^*(A \mid W)$ and the correctly specified case-control weighted targeted maximum likelihood were excluded since those values were the same as those for the targeted maximum likelihood with correctly specified $\hat{Q}^*(A, W)$ and $\hat{g}^*(A \mid W)$.)

as observed in the bias plot displayed in Figure 2(a). The case-control weighted logistic regression and case-control weighted targeted maximum likelihood with misspecified $\hat{Q}^*(A, W)$ had more bias than their correctly specified counterparts. It may be possible to avoid some of the additional bias caused by the misspecification of $\hat{Q}^*(A, W)$ in practice by fitting $\hat{Q}^*(A, W)$ with data-adaptive model-selection, such as the Deletion/Substitution/Addition (DSA) algorithm or other readily available machine learning algorithms. For more details about this procedure we refer to Sinisi and van der Laan (2004). The bias results for the relative risk and risk difference followed similar trends, as can be seen in Figure 2(b) and 2(c). While the case-control weighted logistic regression has low variance when misspecified, it may be more biased than its targeted counterpart. These results bolster our theoretical arguments that gains in efficiency and reduction in bias can be obtained by having a known prevalence probability and using a targeted estimator. Additionally, under typical circumstances experienced in an experimental setting, the case-control weighted targeted maximum likelihood may perform the best with regard to bias and variability.

5.2 Simulation 2

Our second set of simulations was based on a population of $N = 80,000$ individuals, and was designed to illustrate, in another setting, the advantages of incorporating known prevalence probability into case-control design methodology. The population was generated with binary exposure A and disease status Y and a 1-dimensional covariate W. These variables were generated according to the following rules:

$$W \sim U(0,1)$$

$$g_0^*(A \mid W) = P_0^*(A = 1|W) = \frac{1}{1+\exp(-5\sin(W))}$$

$$Q_0^*(A, W) = P_0^*(Y = 1|A, W) = \frac{1}{1+\exp(-(2A-25W+AW))}.$$

The resulting population had a prevalence probability $q_0 = 0.053$, exactly $4,206$ cases, and also followed an independent case-control sampling design. The true values of the risk difference, relative risk, and odds ratio were given by $RD = 0.061$, $RR = 3.21$, and $OR = 3.42$, with $P(Y_1 = 1) = 0.089$ and $P(Y_0 = 1) = 0.028$. These parameters were estimated using the same general methods as in the previous section, albeit with different fits for $\hat{Q}^*(A, W)$ and $\hat{g}^*(A \mid W)$. The initial fit for each method requiring a fit for $\hat{Q}^*(A, W)$ was

Table 3: **Simulation 2 – Odds Ratio** – MSE is Mean Squared Error for IPTW misspecified Estimate, RE is Relative Efficiency of Other Estimators Compared to IPTW misspecified Estimate MSE, nC is Number of Cases, nCo is Number of Controls, n is Number of Total Observations, M is for Misspecified $\hat{Q}^*(A,W)$ or $\hat{g}^*(A \mid W)$ Fit, C is for Correctly Specified $\hat{Q}^*(A,W)$ or $\hat{g}^*(A \mid W)$. (When two letters are noted in the "Fit" column, the first letter refers to $\hat{Q}^*(A,W)$ and the second to $\hat{g}^*(A \mid W)$.)

		n=350	n=500	n=750	n=1000
Odds Ratio	Fit	nC=100 nCo=250	nC=250 nCo=250	nC=250 nCo=500	nC=500 nCo=500
IPTW MSE	M	404.40	3667.56	306.42	2433.62
IPTW RE	C	1.0E+00	1.2E+00	1.0E+00	1.2E+00
CCW-TMLE RE	C/C	2.8E+02	4.1E+03	5.7E+02	5.7E+03
	C/M	2.9E+02	4.1E+03	5.7E+02	5.7E+03
CCW-MLE RE	C	2.9E+02	4.2E+03	5.7E+02	5.8E+03

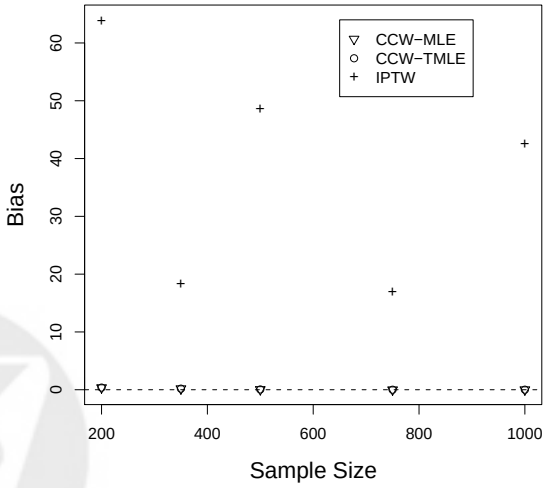


Figure 3: **Simulation 2 – Bias Results for the Odds Ratio.** (Bias results for the case-control weighted targeted maximum likelihood with misspecified $\hat{g}^*(A \mid W)$ were excluded since those values were the same as those for the targeted maximum likelihood with correctly specified $\hat{Q}^*(A,W)$ and $\hat{g}^*(A \mid W)$.)

defined by:

$$\hat{Q}^*(A,W) = \frac{1}{1+\exp(-(\hat{\alpha_0}+\hat{\alpha_1}A+\hat{\alpha_2}W+\hat{\alpha_3}AW))},$$

which was the correctly specified fit. For methods requiring a fit for exposure mechanism, the correct fit was defined by:

$$\hat{g}^*(A \mid W) = \frac{1}{1+\exp(-(\hat{\eta_0}+\hat{\eta_1}\sin(W)))}.$$

The misspecified version of the exposure mechanism was given by:

$$\hat{g}^*(A \mid W) = \frac{1}{1+\exp(-(\hat{\eta_0}+\hat{\eta_1}W))}.$$

Results across the two case-control weighted methods for the risk difference, relative risk, and odds ratio were nearly identical, indicating in this example that when $\hat{Q}^*(A,W)$ is correct and q_0 is known, one may be well served by either of these methods. However, the IPTW method for odds ratio estimation was quite inefficient in comparison. We theorized in van der Laan (2008), and Mansson et al. (2007) demonstrated, that the IPTW procedure has a strong sensitivity towards model misspecification. This result was seen in Simulation 1, although the results in Simulation 2 are more extreme. Results for the odds ratio estimation can be seen in Table 3 and Figure 3. Again we see that gains in efficiency and reduction in bias can be obtained by simply having known q_0.

5.3 Standard Errors, Confidence Intervals, and P-Values

Continuing with the simulated population from Simulation 2, we provide an example of the use of influence curves in the estimation of standard errors for case-control weighted targeted maximum likelihood estimation. We sampled one data set of $n = 1000$ from the population, with equal numbers of cases and controls, and estimated the odds ratio. Recall that the true value for the odds ratio was given by $OR = 3.42$. The case-control weighted targeted maximum likelihood estimator uses the influence curve to estimate standard errors, as discussed in Section 3.2.6, with estimated variance given by $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} D^2_{q_0}(\psi^*, g^*, Q^*)(O)$. Standard error estimates for the IPTW estimator were calculated by bootstrapping the case and control samples 1000 times. The results for this single sampling of the simulated population can be seen in Table 4, including odds ratio estimates, standard errors, confidence intervals, and p-values. It compares only the case-control weighted targeted

Table 4: **Standard Error Illustration** – OR is Odds Ratio Estimate, SE is Standard Error, CI is Confidence Interval, P is P-value, C is for Correctly Specified $\hat{Q}^*(A,W)$ or $\hat{g}^*(A \mid W)$, M is for Misspecified $\hat{g}^*(A \mid W)$. (When two letters are noted in the "Fit" column, the first letter refers to $\hat{Q}^*(A,W)$ and the second to $\hat{g}^*(A \mid W)$.) The results are for one data set of 1000 individuals with 500 cases and 500 controls randomly sampled from the population in Simulation 2. True $OR = 3.42$.

Odds Ratio	Fit	OR	SE	CI	P
IPTW	C	64.98	22.44	[21.00, 108.96]	0.004
	M	64.64	4.66	[55.50, 73.77]	< 0.001
CCW-TMLE RE	C/C	3.39	0.24	[2.93, 3.85]	< 0.001
	C/M	3.39	0.24	[2.92, 3.86]	< 0.001

maximum likelihood estimator and the IPTW estimator. (The non-targeted maximum likelihood method was excluded as we wish to draw attention to the use of the influence curve for standard error estimation. Standard errors for the non-targeted maximum likelihood method can also be calculated using bootstrapping.)

5.4 Simulation 3

Our third simulation study was designed to illustrate the performance of the case-control weighting scheme for targeted maximum likelihood estimation in case-control designs when q_0 is estimated. We also examine coverage probabilities and percentage of rejected tests for case-control weighted targeted maximum likelihood estimation. The simulation was based on a population of $N = 120,000$ individuals, and we simulated a 1-dimensional covariate W, binary exposure A, and indicator Y. The variables were generated according to the following rules:

$$W \sim U(0,1)$$

$$g_0^*(A \mid W) = P_0^*(A = 1|W) = \frac{1}{1+\exp(-(W^2-4W+1))}$$

$$Q_0^*(A,W) = P_0^*(Y = 1|A,W) = \frac{1}{1+\exp(-(A-\sin(W^2)+A\sin(W^2)+7A\log(W)+5\log(W)-1))}.$$

The resulting population had a prevalence probability $q_0 = 0.032$, and exactly $3,834$ cases. We ran 1000 simulations and sampled 500 cases and 500 controls for varying levels of the prevalence probability $q_0 = (0.02, 0.03, 0.04)$. The true

value for the odds ratio was given by $OR = 1.851$, with $P(Y_1 = 1) = 0.052$ and $P(Y_0 = 1) = 0.029$. The causal odds ratio was estimated using case-control weighted targeted maximum likelihood estimation. The correctly specified initial fit for $Q_0^*(A, W)$ was estimated by:

$$\hat{Q}^*(A, W) = \frac{1}{1+\exp(-(\hat{\alpha_0}+\hat{\alpha_1}A+\hat{\alpha_2}\log(W)+\hat{\alpha_3}\sin(W^2)+\hat{\alpha_4}A\log(W)+\hat{\alpha_5}A\sin(W^2)))}.$$

The misspecified initial fit was estimated with:

$$\hat{Q}^*(A, W) = \frac{1}{1+\exp(-(\hat{\alpha_0}+\hat{\alpha_1}A+\hat{\alpha_2}W))}.$$

For exposure mechanism, the correct fit was defined by:

$$\hat{g}^*(A \mid W) = \frac{1}{1+\exp(\hat{\eta_0}+\hat{\eta_1}W^2+\hat{\eta_2}W)}.$$

The misspecified version of the exposure mechanism was given by:

$$\hat{g}^*(A \mid W) = \frac{1}{1+\exp(\hat{\eta_0}+\hat{\eta_1}W)}.$$

When examining the mean squared error results of the odds ratio across the range of values for q_0, one can see deviations away from the values obtained for the true q_0. However, it is important to note that the coverage probabilities (the percentage of simulations where the estimated confidence interval contained the true odds ratio) were not highly variant and remain near 95%. This provides evidence that the case-control weighted targeted maximum likelihood procedure performs well with estimated values of q_0. The percentage of rejected tests ($\alpha = 0.05$) across the range of q_0 was also relatively stable. The results for the mean squared errors, coverage probabilities, and percent rejected tests for the odds ratio can be seen in Table 5. Simulations that resample q_0 from its sampling distribution could also be used to get an estimate of the total uncertainty surrounding the parameter of interest, but they are not explored here. An analytic equivalent to this resampling can be found in the appendix to van der Laan (2008). This theorem demonstrates that one can incorporate the standard error of the estimate $\hat{q}_0$ into the confidence interval for the parameter of interest.

Table 5: **Simulation 3 – Odds Ratio** – MSE is Mean Squared Error, CP is Coverage Probability (percentage of simulations where estimated confidence interval contained the true odds ratio), Rej is for Percent Rejected Tests ($\alpha = 0.05$), C is for Correctly Specified $\hat{Q}^*(A, W)$ or $\hat{g}^*(A \mid W)$, M is for Misspecified $\hat{Q}^*(A, W)$ or $\hat{g}^*(A \mid W)$. (When two letters are noted in the "Fit" column, the first letter refers to $\hat{Q}^*(A, W)$ and the second to $\hat{g}^*(A \mid W)$.) The results are for 1000 simulations of 1000 individuals with 500 cases and 500 controls randomly sampled from the population in Simulation 3. True $OR = 1.851$. True $q_0 = 0.032$.

		True q_0	q_0		
	Fit	0.032	0.020	0.030	0.040
CCW-TMLE MSE	C/C	0.35	0.74	0.39	0.24
	C/M	0.35	0.74	0.39	0.24
	M/C	0.19	0.28	0.20	0.16
CCW-TMLE CP	C/C	0.94	0.95	0.94	0.92
	C/M	0.97	0.97	0.97	0.95
	M/C	0.92	0.94	0.93	0.91
CCW-TMLE Rej	C/C	0.33	0.32	0.33	0.34
	C/M	0.21	0.23	0.22	0.20
	M/C	0.02	0.01	0.02	0.03

6 Discussion

Case-control weighted targeted maximum likelihood estimation provides a framework for the analysis of case-control study designs. We observed that the IPTW method for causal parameter estimation was outperformed in conditions similar to a practical setting by the new case-control weighted targeted maximum likelihood estimation methodology. The case-control weigted targeted maximum likelihood estimation procedure yields a fully robust and locally efficient estimator of several marginal causal parameters of interest. Model misspecification within this framework, with known exposure mechanism, still results in efficient estimatiors. Additionally, the case-control weighted logistic regression mapped to causal parameters had high efficiency and reduced bias in comparison to the IPTW estimator. This is an important result for those applied researchers who may not feel comfortable implementing the case-control weighted targeted maximum likelihood procedure. Thus, we showed striking improvements in efficiency and bias in all methods incorporating knowledge of the prevalence probability over the IPTW estimator which does not use this information. Knowledge of the prevalence probability may be realistic in

many settings. Where possible, researchers might consider prioritizing accurately defining their population of interest, which will streamline obtaining or estimating the prevalence probability. We also demonstrated that a range of values for q_0 can be used with case-control weighted targeted maximum likelihood estimation to obtain efficient causal parameters of interest. As addressed earlier, we discussed case-control weighted targeted maximum likelihood estimation for cumulative study designs with the prevalence probability. Future areas of work include adapting our methods for density sampling, where controls are drawn from the population at risk at the time a case develops disease. For example, using case-control weights that depend on the time points the cases and controls were sampled, as discussed in an appendix in van der Laan (2008). Here, the use of incidence probabilities would be more appropriate.

References

J.A. Anderson. Separate sample logistic discrimination. *Biometrika*, 59:19–35, 1972.

O. Bembom, M.L. Peterson, S-Y Rhee, W.J. Fessel, S.E. Sinisi, R.W. Shafer, and M.J. van der Laan. Biomarker discovery using targeted maximum likelihood estimation: Application to the treatment of antiretroviral resistant hiv infection. *Technical Report 221, Division of Biostatistics, University of California, Berkeley*, 2007.

J. Benichou and S. Wacholder. A comparison of three approaches to estimate exposure-specific incidence rates from population-based case-control data. *Statistics in Medicine*, 13:651–661, 1994.

N.E. Breslow. Statistics in epidemiology: the case-control study. *J Am Stat Soc*, 91:14–28, 1996.

N.E. Breslow and N.E. Day. *Statistical Methods in Cancer Research: Volume 1 – The analysis of case-control studies*. International Agency for Research on Cancer, Lyon, 1980.

J. Cornfield. A method of estimating comparative rates from clinical data. applications to cancer of the lung, breast, and cervix. *J Nat Cancer Inst*, 11:1269–1275, 1951.

J. Cornfield. A statistical problem arising from retrospective studies. In J. Neyman, editor, *Proceedings of the Third Berkeley Symposium, Volume IV*, pages 133–148. University of California Press, 1956.

S. Greenland. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *Am J Epidemiol*, 160(4):301–305, 2004.

S. Greenland. Multivariate estimation of exposure-specific incidence from case-control studies. *J Chron Dis*, 34:445–453, 1981.

P.W. Holland and D.B. Rubin. Causal inference in retrospective studies. In D.B. Rubin, editor, *Matched Sampling for Causal Effects.* Cambridge University Press, Cambridge, MA, 1988.

R. Mansson, M.M. Joffe, W Sun, and S. Hennessy. On the estimation and use of propensity scores in case-control and case-cohort studies. *Am J Epidemiol*, 166(3):332–339, 2007.

K.L. Moore and M.J. van der Laan. Covariate adjustment in randomized trials with binary outcomes. *Technical Report 215, Division of Biostatistics, University of California, Berkeley*, 2007.

A.P. Morise, G.A. Diamon, R. Detrano, M. Bobbio, and Erdogan Gunel. The effect of disease-prevalence adjustments on the accuracy of a logistic prediction model. *Med Decis Making*, 16:133–142, 1996.

S. Newman. Causal analysis of case-control data. *Epid Persp Innov*, 3:2, 2006. URL `http://www.epi-perspectives.com/content/3/1/2`.

R.L. Prentice and N.E. Breslow. Retrospective studies and failure time models. *Biometrika*, 65(1):153–158, 1978.

R.L. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66:403–411, 1979.

J.M. Robins. [choice as an alternative to control in observational studies]: Comment. *Statistical Science*, 14(3):281–293, 1999.

S. Rose and M.J. van der Laan. Why match? investigating matched case-control study designs with causal effect estimation. *Technical Report 240, Division of Biostatistics, University of California, Berkeley*, 2008.

K. Rothman and S. Greenland. *Modern Epidemiology.* Lippincott, Williams and Wilkins, Philadelphia, PA, 2nd edition, 1998.

S. Sinisi and M.J. van der Laan. Deletion/substitution/addition algorithm in loss function based estimation. *Journal of Statistical Methods in Molecular Biology*, 3(1):Article 18, 2004.

M.J. van der Laan. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1):Article 2, 2006.

M.J. van der Laan. Estimation based on case-control designs with known incidence probability. *The International Journal of Biostatistics*, 4(1):Article 17, 2008.

M.J. van der Laan and J.M. Robins. *Unified methods for censored longitudinal data and causailty*. Springer, New York, 2002.

M.J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):Article 11, 2006.

S. Wacholder. The case-control study as data missing by design: Estimating risk differences. *Epidemiology*, 7(2):144–150, 1996.

8.3 *Why Match? Investigating Matched Case-Control Study Designs with Causal Effect Estimation*

The following article appears as it was published in the *International Journal of Biostatistics* in 2009, http://www.bepress.com/ijb/vol5/iss1/1/.

It was originally published on the University of California, Berkeley Division of Biostatistics Working Paper Series website in 2008, http://www.bepress.com/ucbbiostat/paper240/.

Why Match? Investigating Matched Case-Control Study Designs with Causal Effect Estimation

Sherri Rose and Mark J. van der Laan
Division of Biostatistics, University of California, Berkeley
sherri@berkeley.edu, laan@berkeley.edu

Abstract

Matched case-control study designs are commonly implemented in the field of public health. While matching is intended to eliminate confounding, the main potential benefit of matching in case-control studies is a gain in efficiency. Methods for analyzing matched case-control studies have focused on utilizing conditional logistic regression models that provide conditional and not causal estimates of the odds ratio. This article investigates the use of case-control weighted targeted maximum likelihood estimation to obtain marginal causal effects in matched case-control study designs. We compare the use of case-control weighted targeted maximum likelihood estimation in matched and unmatched designs in an effort to explore which design yields the most information about the marginal causal effect. The procedures require knowledge of certain prevalence probabilities and were previously described by van der Laan (2008). In many practical situations where a causal effect is the parameter of interest, researchers may be better served using an unmatched design.

1 Introduction

Individually matched case-control study designs are frequently found in public health and medical literature, and conditional logistic regression is the tool most commonly used to analyze these studies. Matching is intended to eliminate confounding, however, the main potential benefit of matching in case-control studies is a gain in efficiency. Therefore, when are these study designs truly beneficial? Given all the potential drawbacks, including extra cost, added time for enrollment, and increased bias, the use of matching in case-control study designs warrants careful evaluation. Discussion of the advantages and disadvantages of matching in the literature goes back more than 40 years.

In this paper, we focus on individual matching in case-control studies where the researcher is interested in estimating the *marginal causal* effect, and certain prevalence probabilities are known. Our procedure, first presented in van der Laan (2008), "targets" the parameter of interest rather than the distribution of interest, and is thus aptly named case-control weighted targeted maximum likelihood estimation. In order to eliminate the bias caused by the matched case-control sampling design, this technique relies on knowledge of the true prevalence probability $q_0 \equiv P_0^*(Y = 1)$, and an additional value $\bar{q}_0(M) \equiv q_0 \frac{P_0^*(Y=0|M)}{P_0^*(Y=1|M)}$, where M is the matching variable. For unmatched designs, knowledge of only q_0 is required.

The case-control weighting scheme maps estimation methods developed for prospective sampling into methods for case-control sampling, and it produces efficient estimators when its prospective sample counterpart is efficient. Thus, both the matched and unmatched procedures are double robust and locally efficient: they perform well as long as $P_0^*(Y \mid A, W)$ or $P_0^*(A \mid W)$ is correctly specified, are consistent if either of these models are correctly specified, and efficient if both are correctly specified. (Here A is the exposure of interest and W is a vector of covariates.) We will compare the use of case-control weighted targeted maximum likelihood estimation in matched and unmatched case-control study designs as we explore which design yields the most information about the marginal causal effect. This paper will not address matching in cohort studies, and will concentrate solely on case-control studies. However, matching in cohort studies was briefly addressed in van der Laan (2008), and applying our methods to cohort studies is an area of future research.

2 Why Match? A Literature Review

There is a large collection of literature devoted to the topic of individual matching in case-control study designs. This overview attempts to capture the most important consideratons, and it is by no means exhaustive.

2.1 Individual Matching in Case-Control Studies

In an individually matched case-control study, the population of interest is identified, and cases are randomly sampled or selected based on particular inclusion criteria. Although, as Rothman and Greenland (1998) note, the definition of a case may implicitly define the population of interest for cases and controls. Each of these cases is then matched to one or more controls based on a variable (or variables) *believed* to be a confounder. Much of the literature on individual matching in case-control studies, particularly earlier texts, describes these designs as a way to reduce confounding in the sampling design. Reference to this is made in: Miettinen (1970), Breslow et al. (1978), Breslow and Day (1980), Kupper et al. (1981), Schlesselman (1982), Collett (1991), and Costanza (1995), among others. However, several authors (Breslow and Day, 1980; Kupper et al., 1981; Schlesselman, 1982; Rothman and Greenland, 1998; Vandenbroucke et al., 2007) point out that the goal of matching is to increase the study's efficiency by forcing the case and control samples to have similar distributions across confounding variables. Rothman and Greenland (1998) go on to say that while matching is intended to control confounding, it cannot do this in case-control study designs, and can, in fact, introduce bias. Costanza (1995) agreed, stating that matching on confounders in case-control studies does nothing to remove the confounding, but frequently introduces negative confounding.

So, while some literature cites the purpose of matching as improving validity, later publications (Kupper et al., 1981; Rothman and Greenland, 1998) demonstrated that matching has a greater impact on efficiency over validity. Matched sampling leads to a balanced number of cases and controls across the levels of the selected matching variables. This balance can reduce the variance in the parameters of interest, which improves statistical efficiency. A study with a randomly selected control group may yield some strata with an imbalance of cases and controls. It is important to add, however, that matching in case-control studies can lead to gains *or* losses in efficiency (Kupper et al., 1981; Rothman and Greenland, 1998). This will be discussed further in later sections.

Breslow and Day (1980) note that matched case-control studies attempt to increase the informativeness of each of the subjects in the study. However, one should also note that matched studies discard not only a pool of unmatched controls, but the information in each exposure-concordant case-control pair. Additionally, matching has a substantial impact on the study sample, most notably, it creates a sample of controls that is not representative of exposure in the population or the population as a whole. The effect of the matching variable can no longer be studied directly, and the exposure frequency in the control sample will be shifted towards that of the cases (Rothman and Greenland, 1998). Matching in case-control studies also does not completely control for the variable or variables used for matching, in general. This means that researchers who implement matched designs must perform matched or stratified analyses (Seigel and Greenhouse, 1973; Schlesselman, 1982; Holland and Rubin, 1988; Rothman and Greenland, 1998; Rubin, 2006). If an unmatched analysis is performed on matched data, the validity of the case-control comparison may be decreased (Schlesselman, 1982).

2.2 Variable Selection

We revisit an earlier point made in this overview of individually matched case-control studies: matching variables are chosen *a priori* on the belief that they confound the relationship between exposure and disease. If controls are matched to cases based on a variable that is not a true confounder, this can impact efficiency. For example, if the matching variable is not associated with disease but is associated with the exposure, this will increase the variance of the estimator compared to an unmatched design. Here, the matching leads to larger numbers of exposure-concordant case-control pairs, which are not informative in the analysis, leading to the increase in variance. If the matching variable is only associated with disease, there is often a loss of efficiency as well (Schlesselman, 1982). If the matching variable is along the causal pathway between disease and exposure then matching will contribute bias that cannot be removed in the analysis (Vandenbroucke et al., 2007). Matching on a variable associated with exposure and not disease or a variable along the causal pathway are considered types of *overmatching*. Variables for matching should therefore be selected very carefully, and only those that are known to be associated with both exposure and disease should be considered. The number of matching variables should also be reduced to as few as possible. As the number of matching variables grows, the cases and controls will become increasingly similar with respect to the exposure of interest, and the study may produce a spurious result or provide no information (Breslow and Day, 1980).

Additionally, when matching on more than one variable, matching variables should not be strongly correlated with each other (Schlesselman, 1982).

2.3 More on Efficiency

Kupper et al. (1981) performed a variety of simulations to demonstrate the impact of matching on efficiency. They found that in situations where confounding was present, the confidence intervals for matched studies were smaller than unmatched studies unless the odds ratio and the exposure of interest were large. However, the confidence intervals for the samples with randomly selected controls were always shorter when the number of controls was at least twice that of the cases. This is an important result, as efficiency is often touted as the benefit of an individually matched case-control study design. Simulations aside, Cochran (1953) is often cited as the theoretical paper that demonstrates the efficiency of matched designs. However, as noted by McKinlay (1977), Cochran's result can be misleading. Comparisons between matched and unmatched study designs are often made with *equal* sample sizes and no other method of covariate adjustment (e.g. regression). In a matched design, controls may be discarded if they do not match a particular case on the variable or variables of interest. Multiple controls may be discarded per case, depending on the variables of interest (Freedman, 1950; Cochran and Chambers, 1965; McKinlay, 1977). In a typical randomly selected case-control study, these controls would be included. In many cases, if the discarded controls were available to be rejected in the matched study, they would be available for an unmatched design in the same investigation (Billewicz, 1965; McKinlay, 1977). Therefore, it may be more appropriate to compare the efficiencies of matched case-control studies of size n to randomly selected case-control studies of size n+*number of discarded controls*. Additionally, these randomly selected case-control studies should employ a method of analysis to reduce bias and variance. Therefore, the result from Kupper et al. (1981) is especially poignant, as all randomly selected case-control studies that had a size of at least $2n$ had shorter confidence intervals than their matched counterparts of size n.

2.4 Trends

Gefeller et al. (1998) performed a literature review of case-control studies published between 1955 and 1994 in three main epidemiology journals: *American Journal of Epidemiology, International Journal of Epidemiology, and the Journal of Epidemiology and Community Health*. They found that, among these journals, there was a decreasing trend in the percentage of individually

matched case-control studies published (71.7% in the years preceding 1981, 65.5% in 1985, 46.9% in 1989, and 46.4% in 1994), and an increasing percentage of frequency matched studies (5.0% in the years preceding 1981, 9.1% in 1985, 16.3% in 1989, and 26.2% in 1994). Interestingly, the percentage of case-control studies using no matching stayed relatively constant with no obvious trend (averaging 29.3%, and ranging from 23.2% to 36.7%). Unfortunately, they found substantial evidence that individually matched studies were being performed without the appropriate matched analysis: only 74% of studies from 1994 used conditional logistic regression if logistic regression was the chosen method of analysis. A later analysis of medical literature in Medline, Rahman (2003), indicated that 5.3% of individually matched case-control studies used an unconditional logistic regression for those selecting logistic regression models. The review in Gefeller et al. (1998) indicates that unmatched case-control studies, at least in epidemiology, are in the minority. This should be questioned given the overwhelming agreement in the literature that matching is not frequently justified for case-control study designs.

2.5 Literature Review Discussion

The consensus in the literature indicates that there are very few circumstances where individual matching is indeed warranted. Case-control studies with a very small number of cases may benefit from individual matching, as a randomly selected control group from even a well-defined population of interest may be uninformative on many variables of interest (Schlesselman, 1982; Costanza, 1995). Individual matching moves from beneficial to required when variables such as sibship are included in the study (Rothman and Greenland, 1998; Costanza, 1995). Matching is also cited as necessary by many authors when the investigators expect the distribution of the matching variable to differ drastically between the cases and the controls. It may be this reason that draws many investigators towards a matched design, perhaps without appropriate consideration of the disadvantages or definition of the population of interest.

Methodologists in the literature stress that it is often possible for confounders to be *adjusted for* in the analysis instead of matched on in case-control designs (Schlesselman, 1982; Vandenbroucke et al., 2007). The development of effective methods to control confounding in analyses may have contributed to the drop in individually matched designs, but they are still quite common. It is therefore important to continue to disseminate the implications of individually matched case-control study designs to researchers, as Rothman and Greenland (1998) note that *"people match on a variable (e.g. sex) simply because it is*

the 'expected thing to do' and they might lose credibility for not matching." When researchers make design and analysis decisions based on these types of considerations, their research may suffer.

Our contributions to the vast literature on individual matching for case-control studies will be unique. We focus on scenarios where the researcher is interested in estimating a marginal causal effect, a parameter that cannot be estimated with conditional logistic regression, and certain prevalence probabilities are known. Thus, we will compare the use of case-control weighted targeted maximum likelihood estimation in matched and unmatched designs in an effort to explore which design yields the most information about the marginal causal effect.

3 Existing Methods

Model-based methods for the analysis of matched case-control studies are plentiful in recent literature (Breslow et al., 1978; Holford et al., 1978; Breslow and Day, 1980; Greenland, 1981; Schlesselman, 1982; Holland and Rubin, 1988; Benichou and Wacholder, 1994; Rothman and Greenland, 1998; Greenland, 2004). And, while it is not the only method of analysis for individually matched case-control studies, the predominant method of analysis is conditional logistic regression. This method provides a conditional estimate of the odds ratio of being diseased given the exposure of interest and baseline covariates. Conditional logistic regression will be discussed in more detail in the subsection below. Greenland (1981) and Holland and Rubin (1988) discuss another model-based method: the use of log-linear models to estimate the marginal odds ratio. Additionally, Rothman and Greenland (1998) and Greenland (2004) demonstrate the use of standardization in case-control studies, which estimate marginal effects with population or person-time averaging. Holland and Rubin (1988) note that the traditional two-way table and its extensions generally provide no causal insight for matched case-control studies. However, these methods are all distinctly different from the method we illustrate in this paper, discussed by van der Laan (2008), as our method is a nonparametric double robust locally efficient procedure that provides an estimate of the marginal causal odds ratio.

3.1 Conditional Logistic Regression

The logistic regression model for matched case-control studies differs from unmatched studies in that it allows the intercept to vary among the matched

units of cases and controls. The matching variable is not included in the model (Breslow et al., 1978; Holford et al., 1978; Breslow and Day, 1980; Schlesselman, 1982). If the parameter of interest is the coefficient in front of the exposure A, the use of a matched study design and a conditional logistic regression analysis can yield increases in efficiency, compared to an unmatched design with a logistic regression analysis. It is important to note that in order to estimate an effect of exposure A with conditional logistic regression, the case and control must be discordant on A. Additionally, if information for a variable is missing for a case (or control), the corresponding control (or case) information is discarded (Breslow and Day, 1980; Schlesselman, 1982). These two limitations do not occur in the new case-control weighted targeted maximum likelihood estimation methodology for causal effect parameters. More importantly, if a marginal causal effect is the parameter of interest, conditional logistic regression cannot be used as it can only estimate the conditional odds ratio.

4 Case-Control Weighted Targeted Maximum Likelihood Estimation

4.1 Background

We define $O^* = (W, A, Y) \sim P_0^*$ as the experimental unit and corresponding distribution P_0^* of interest. P_0^* represents the population from which all cases and controls will be sampled. Here O^* consists of baseline covariates W, an exposure variable A (referred to as the "treatment" variable in prospective studies), and a binary outcome Y, which defines case or control status. If we are interested in marginal causal effect parameters, we can define $\psi_0^* = \Psi^*(P_0^*) \in \mathbb{R}^d$ of $P_0^* \in \mathcal{M}^*$ as the causal effect parameter and define the risk difference, relative risk, odds ratio as follows for binary exposure $A \in \{0, 1\}$:

$$\begin{aligned} \psi_{0,RD}^* &\equiv E_0^*\{E_0^*(Y \mid A = 1, W) - E_0^*(Y \mid A = 0, W)\} \\ &= E_0^*(Y_1) - E_0^*(Y_0) \\ &= P_0^*(Y_1 = 1) - P_0^*(Y_0 = 1), \end{aligned} \tag{1}$$

$$\psi_{0,RR}^* = \frac{E_0^* E_0^*(Y \mid A = 1, W)}{E_0^* E_0^*(Y \mid A = 0, W)} = \frac{E_0^*(Y_1)}{E_0^*(Y_0)} = \frac{P_0^*(Y_1 = 1)}{P_0^*(Y_0 = 1)}, \tag{2}$$

and,

$$\psi_{0,OR}^* = \frac{P_0^*(Y_1 = 1) P_0^*(Y_0 = 0)}{P_0^*(Y_1 = 0) P_0^*(Y_0 = 1)}. \tag{3}$$

These causal versions of the effect parameters require the specification of the counterfactual outcomes Y_0 and Y_1 for binary A and $(W, A, Y = Y_A)$ as a time-ordered missing data structure on the full data structure (W, Y_0, Y_1). One must also make the randomization assumption: $\{A \perp Y_0, Y_1 \mid W\}$. Since these parameters are always well defined parameters of the distribution of the data, they can thereby be viewed as W-adjusted variable importance parameters. Then there is no need to make these assumptions. We refer to van der Laan (2006) for the details of this framework.

However, the observed data structure in matched case-control sampling is defined by:

$$O = ((M_1, W_1, A_1), (M_0^j = M_1, W_0^j, A_0^j : j = 1, \ldots, J)) \sim P_0, \text{ with}$$

$$(M_1, W_1, A_1) \sim (M, W, A \mid Y = 1) \text{ for cases, and}$$

$$(M_0^j, W_0^j, A_0^j) \sim (M, W, A \mid Y = 0, M = M_1) \text{ for controls.}$$

Here $M \subset W$, and M is a categorical matching variable. The sampling distribution of the data structure O is described as above with P_0. Thus, the matched case-control data set contains n independent and identically distributed observations $O_1, \ldots, O_n$ with sampling distribution P_0. The cluster containing one case and the J controls is the experimental unit, and the marginal distribution of the cluster is specified by the population distribution P_0^*. The model $\mathcal{M}^*$, which possibly includes knowledge of q_0 or $\bar{q}_0(M)$, then implies models for the marginal distribution of cases (M_1, W_1, A_1) and controls $(M_1, W_2^j, A_2^j), j = 1, \ldots, J$.

Independent case-control sampling is described as sampling nC cases from the conditional distribution of (W, A), given $Y = 1$, and sampling nCo controls from (W, A), given $Y = 0$. The value of J used to weight each control is then nCo/nC. We refer to independent case-control sampling as Case-Control Design I, and matched case-control sampling as Case-Control Design II.

4.2 Methodology Summary

If one wishes to estimate marginal causal effects for Case-Control Design II, which correspond with the traditional parameters of interest in randomized trials, there is now a nonparametric double robust locally efficient procedure available. It performs well as long as $P_0^*(Y \mid A, W)$ or $P_0^*(A \mid W)$ is correctly specified, is consistent if either of these models are correctly specified, and efficient if both are correctly specified. The theoretical framework for case-control weighted targeted maximum likelihood estimation has been discussed

in detail in van der Laan (2008), and step-by-step implementation for Case-Control Design I appears in Rose and van der Laan (2008). For the targeted maximum likelihood framework designed for prospective sampling, see van der Laan (2006), and for its implementation, see Bembom et al. (2007).

Case-control weighted targeted maximum likelihood estimation for Case-Control Design II incorporates estimates of $P_0^*(Y \mid A, W)$, $P_0^*(A \mid W)$, and knowledge of q_0 and $\bar{q}_0(M)$, where $\bar{q}_0(M)$ is defined as:

$$\bar{q}_0(M) \equiv q_0 \frac{P_0^*(Y = 0 \mid M)}{P_0^*(Y = 1 \mid M)} = q_0 \frac{q_0(0 \mid M)}{q_0(1 \mid M)}.$$

The case-control weighted targeted maximum likelihood estimation procedure for Case-Control Design II uses $P_0^*(A \mid W)$ to update an initial estimate of $P_0^*(Y \mid A, W)$.

4.3 Implementation

Case-control weighted targeted maximum likelihood estimation for Case-Control Designs I and II can be implemented using existing software (including `SAS`, `STATA`, and `R`). The implementation of case-control weighted targeted maximum likelihood for Case-Control Design II is also very similar to the implementation for Case-Control Design I. Key differences will be stressed here, but for more detail, we refer to Rose and van der Laan (2008).

Weighting. Weights q_0 and $\bar{q}_0(M)\frac{1}{J}$ are assigned to the cases and corresponding J controls, respectively. *This differs from Case-Control Design I in that* $(1 - q_0)\frac{1}{J}$ *is used to weight controls in Case Control Design I instead of* $\bar{q}_0(M)\frac{1}{J}$. In van der Laan (2008) it is suggested that in cases where $\bar{q}_0(M)$ is not known, $1 - q_0$ can be used to approximate $\bar{q}_0(M)$.

Estimating $Q_0^*(A, W)$. Estimate $P_0^*(Y \mid A, W) \equiv Q_0^*(A, W)$ using the appropriate weights. This estimate is denoted $\hat{Q}^*(A, W)$. Two methods for estimating $\hat{Q}^*(A, W)$ include intercept adjusted logistic regression and case-control weighted logistic regression. Intercept adjusted logistic regression adds the intercept $\log q_0/(1 - q_0)$ to a logistic regression model. This yields the true logistic regression function $P_0^*(Y = 1 \mid A, W)$. *If intercept adjusted logistic regression is used to obtain* $\hat{Q}^*(A, W)$*, cases are weighted 1 and controls are weighted with* $\bar{q}_0(M)\frac{1}{J}$*. This is the only step and method where assigned weights are not* q_0 *and* $\bar{q}_0(M)\frac{1}{J}$. In Rose and van der Laan (2008), we discussed disadvantages associated with using intercept adjusted logistic regression, and

thus our simulations will focus on the use of case-control weighted logistic regression for estimating $Q_0^*(A, W)$.

Case-control weighted logistic regression uses the assigned weights and performs maximum likelihood estimation for prospective sampling (ignoring the case-control sampling design). Consider a nonparametric model for the marginal distribution of the covariates, and a model $\{Q_\theta^* : \theta\}$ for $Q_0^*(A, W)$. Then the case-control weighted maximum likelihood estimator for $Q_0^*(A, W)$ in Case-Control Design II is given by:

$$\hat{\theta} = \arg\max_\theta \sum_{i=1}^{n} q_0 \log \hat{Q}_\theta^*(M_{1i}, W_{1i}, A_{1i}) + \bar{q}_0(M_1)\frac{1}{J}\sum_{j=1}^{J} \log(1 - \hat{Q}_\theta^*(M_{1i}, W_{2i}^j, A_{2i}^j)).$$

If $\hat{Q}^*(A, W)$ is obtained using case-control weighted logistic regression, it is weighted with q_0 and $\bar{q}_0(M)\frac{1}{J}$. For further discussion see van der Laan (2008) and Rose and van der Laan (2008).

Estimating $g_0^*(A \mid W)$. Estimate $P_0^*(A \mid W) \equiv g_0^*(A \mid W)$ using assigned weights. This estimate is denoted $\hat{g}^*(A \mid W)$, and may be obtained using case-control weighted logisitic regression, for example.

Calculating $h(A, W)$. Calculate the "clever covariate" for each subject based on $g_0^*(A \mid W)$. The covariate takes the form:

$$h(A, W) \equiv \left(\frac{I(A = 1)}{\hat{g}^*(A = 1 \mid W)} - \frac{I(A = 0)}{\hat{g}^*(A = 0 \mid W)} \right)$$

for the risk difference. Two covariates are used for estimation of other parameters, such as the odds ratio:

$$h_0(A, W) \equiv \left(- \frac{I(A = 0)}{\hat{g}^*(A = 0 \mid W)} \right) \text{ and } h_1(A, W) \equiv \left(\frac{I(A = 1)}{\hat{g}^*(A = 1 \mid W)} \right)$$

For further discussion see van der Laan and Rubin (2006) and Moore and van der Laan (2007).

Updating $\hat{Q}^*(A, W)$. Update $\hat{Q}^*(A, W)$ by performing an additional weighted regression with $h(A, W)$ as a supplementary covariate. The other coefficients in the initial fit $\hat{Q}^*(A, W)$ are held fixed, and the intercept is suppressed in order to estimate the case-control weighted estimator of ϵ, the coefficient in

front of $h(A,W)$, which we denote as $\hat{\epsilon}^1$. The regression estimate $\hat{Q}^*(A,W)$ is then updated and given by $\hat{Q}^*_1(A,W)$:

$$\hat{Q}^*_1(A,W) = \hat{Q}^*(A,W) + \hat{\epsilon}^1 h(A,W).$$

This step is iterated until convergence, although convergence is often achieved in one step.

Estimating Causal Parameters. Using q_0, $\bar{q}_0(M_1)$, and $\hat{Q}^*_1(A,W)$, estimate causal parameters of interest (risk difference, relative risk, and odds ratio, defined in formulas (1), (2), and (3)) by averaging over the case-control weighted distribution of W. This mapping is performed by evaluating $\hat{Q}^*_1(A,W)$ at $A=1$ and $A=0$ and applying weights q_0 to cases and $\bar{q}_0(M_1)\frac{1}{J}$ to the controls. This forms case-control weighted estimates of $E^*_0(Y_1) = P^*_0(Y_1=1)$ and $E^*_0(Y_0) = P^*_0(Y_0=1)$. The causal parameters of interest can then be calculated from these estimates. For example, the relative risk $E^*_0(Y_1)/E^*_0(Y_0)$ is estimated by:

$$\hat{\psi}_{RR} = \frac{\frac{1}{n}\sum_{i=1}^n q_0 \hat{Q}^*_{1,q_0}(M_1, W_{1i}, 1) + \bar{q}_0(M_1)\frac{1}{J}\sum_j \hat{Q}^*_{1,q_0}(M_1, W^j_{2i}, 1)}{\frac{1}{n}\sum_{i=1}^n q_0 \hat{Q}^*_{1,q_0}(M_1, W_{1i}, 0) + \bar{q}_0(M_1)\frac{1}{J}\sum_j \hat{Q}^*_{1,q_0}(M_1, W^j_{2i}, 0)}.$$

Calculating Standard Errors. Calculating standard errors, p-values, and confidence intervals for case-control weighted targeted maximum likelihood estimates requires the use of the case-control weighted influence curve. This methodology is discussed in detail in van der Laan (2008). We also refer to van der Laan and Robins (2002) for careful discussions of gradients and influence curve theory. The case-control weighted influence curve for matched case-control study designs is the influence curve for prospective targeted maximum likelihood with case-control weighting. We refer to van der Laan and Rubin (2006) and Moore and van der Laan (2007) for this methodology. A complete understanding of the derivation of infuence curves is not required to implement the case-control targeted maximum likelihood estimation procedure for Case-Control Design II.

For illustration, we present the unweighted influence curve for the risk difference of a prospective study $\psi^*_{0,RD} = P^*_0(Y_1=1) - P^*_0(Y_0=1)$, which is estimated by:

$$\begin{aligned}\hat{D}_{RD}(\psi^*, g^*, Q^*)(O) = {} & \frac{I(A=1)}{\hat{g}^*(1 \mid W)}(Y - \hat{Q}^*(1,W)) - \frac{I(A=0)}{\hat{g}^*(0 \mid W)}(Y - \hat{Q}^*(0,W)) \\ & + \hat{Q}^*(1,W) - \hat{Q}^*(0,W) - \hat{\psi}.\end{aligned}$$

The case-control weighted double robust efficient influence curve for the risk difference $\psi^*_{0,RD} = P^*_0(Y_1 = 1) - P^*_0(Y_0 = 1)$ in a matched case-control study design is then:

$$\begin{aligned}\hat{D}_{RD,q_0}(\psi^*, g^*, Q^*)(O) &= q_0 \hat{D}^*(g^*, Q^*)(M_1, W_1, A_1, 1) \\ &\quad + \bar{q}_0(M_1)\frac{1}{J}\sum_{j=1}^{J} \hat{D}^*(g^*, Q^*)(M_1, W_2^j, A_2^j, 0) - \psi^*,\end{aligned}$$

The asymptotic variance of $\sqrt{n}(\hat{\psi} - \psi^*_0)$ using the estimate of the efficient influence curve $D_{q_0}(\psi^*, g^*, Q^*)(O)$ can be estimated by:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} D^2_{q_0}(\psi^*, g^*, Q^*)(O).$$

A 95% Wald-type confidence interval can then be constructed using the causal parameter estimate $\hat{\psi}$: $\hat{\psi} \pm z_{0.975}\frac{\hat{\sigma}}{\sqrt{n}}$, as well as a p-value for $\hat{\psi}$: $2[1-\Phi(|\frac{\hat{\psi}}{\hat{\sigma}/\sqrt{n}}|)]$.

5 Simulation Studies

5.1 Simulation 1

Our first simulation study was designed to illustrate the differences between independent case-control sampling (Case-Control Design I) and matched case-control sampling (Case-Control Design II) using the case-control weighting scheme for targeted maximum likelihood estimation proposed by van der Laan (2008). It was also designed to represent "ideal" situations where control information is not discarded (e.g. data collection is expensive, and covariate information is only collected when a control is a match). This simulation also demonstrates the use of weights q_0 and $(1 - q_0)\frac{1}{J}$ with matched data, to represent situations where $\bar{q}_0(M)$ is not known. The population contained $N = 35,000$ individuals, where we simulated a 9-dimensional covariate $W = (W_i : i = 1, \ldots, 9)$, a binary exposure (or "treatment") A, and an indicator Y, which was 1 for cases and 0 for controls. These variables were generated according to the following rules:

$$P^*_0(W_i = 1) = 0.5$$

$$g^*_0(A \mid W) = \frac{1}{1+\exp(-(W_1+W_2+W_3-2W_4-2W_5+2W_6-4W_7-4W_8+4W_9))}$$

$$Q_0^*(A,W) = \frac{1}{1+\exp(-(1.5A+W_1-2W_2-4W_3-W_4-2W_5-4W_6+W_7-2W_8-4W_9))}.$$

It can be seen in both $g_0^*(A \mid W)$ and $Q_0^*(A,W)$ that the covariates were generated with varied levels of association with A and Y. This was done to investigate the role of weak, medium, and strong association between a matching variable W_i and A and Y. The corresponding associations can be seen in Figure 1. For example, W_1 was weakly associated with both A and Y. One might recall that matching is potentially beneficial only when the matching variable is a true confounder; associated with both A and Y.

Figure 1: **Simulated Covariates**

		Y		
	Association	Weak	Medium	Strong
	Weak	W_1	W_2	W_3
A	Medium	W_4	W_5	W_6
	Strong	W_7	W_8	W_9

Another illustration of the varied association levels can be seen in Figure 2. Here, we display the probability an individual in the population was a case given $W_i = 1$, all the non-matching covariates (Z), and A. Likewise, probabilities for $W_i = 0$ are also shown. For example, let's say matching variable W_2 is *age* with 1 representing 'young' (< 50 years) and 0 representing 'old' (≥ 50 years). In this population, it was not very likely (0.013) that someone who is young will become a case, while someone who is old has a much higher chance of becoming a case (0.047), given Z and A. Therefore, W_2, W_5, and W_8 represent situations where the distribution of W_i among cases and controls is very different. The covariates W_3, W_6, and W_9 represent situations where this difference is even more extreme.

The simulated population had a prevalence probability $q_0 = 0.030$, and exactly 1045 cases. The true value of the odds ratio was given by $OR = 2.302$, with $P_0^*(Y_1 = 1) = 0.055$ and $P_0^*(Y_0 = 1) = 0.025$. We sampled the population using a varying number of cases $nC = (200, 500, 1000)$ for both Case-Control Designs I and II, and for each sample size we ran 1000 simulations. For each simulation, the same sampled cases were used for Case Control Designs I and II. Controls were matched to cases on one variable (W_i) in Case-Control Design II for both 1:1 and 1:2 designs. The same number of controls were used in both Case-Control Designs I and II. Causal effect parameters were estimated using case-control weighted targeted maximum likelihood estimation (CCW T-MLE) for Case-Control Designs I and II with case-control weighted logistic

Figure 2: **Simulated Covariates: Probabilities**. Z represents the remaining eight non-matching covariates.

W_i	$P_0^*(Y=1\|W_i=1,Z,A)$	$P_0^*(Y=1\|W_i=0,Z,A)$
W_1	0.039	0.021
W_2	0.013	0.049
W_3	0.003	0.060
W_4	0.021	0.040
W_5	0.013	0.047
W_6	0.003	0.061
W_7	0.040	0.023
W_8	0.013	0.046
W_9	0.004	0.066

regression for $\hat{Q}^*(A,W)$ discussed in Section 4.3. The initial fit for the estimate of $Q_0^*(A,W)$ was correctly specified as:

$$\hat{Q}^*(A,W) = \frac{1}{1+\exp(-(\hat{\alpha_0}+\hat{\alpha_1}A+\hat{\alpha_2}W_1+\hat{\alpha_3}W_2+...+\hat{\alpha_9}W_8+\hat{\alpha_{10}}W_9))}.$$

The initial fit for the exposure mechanism, which was the correct fit, was defined by:

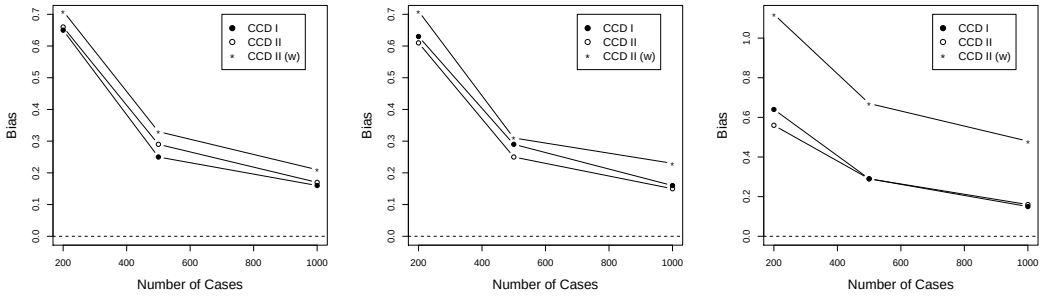
$$\hat{g}^*(A \mid W) = \frac{1}{1+\exp(\hat{\eta_0}+\hat{\eta_1}W_1+\hat{\eta_2}W_2+\hat{\eta_3}W_3+\hat{\eta_4}W_4+\hat{\eta_5}W_5+\hat{\eta_6}W_6+\hat{\eta_7}W_7+\hat{\eta_8}W_8+\hat{\eta_9}W_9)}.$$

Case-Control Designs I and II performed similarly with respect to bias for the nine covariates. When examining efficiency, there were consistent increases in efficiency when the association between W_i and Y was high (W_3, W_6, and W_9), when comparing Case-Control Design II to Case-Control Design I. Results when association with W_i and Y was medium (W_2, W_5, and W_8) were not entirely consistent, although covariates W_5 and W_8 did show increases in efficiency for Case-Control Design II for all or nearly all sample sizes. These results were in line with the consensus found in our literature search: that matching may produce gains in efficiency when the distribution of the matching variable differs drastically between the cases and the controls.

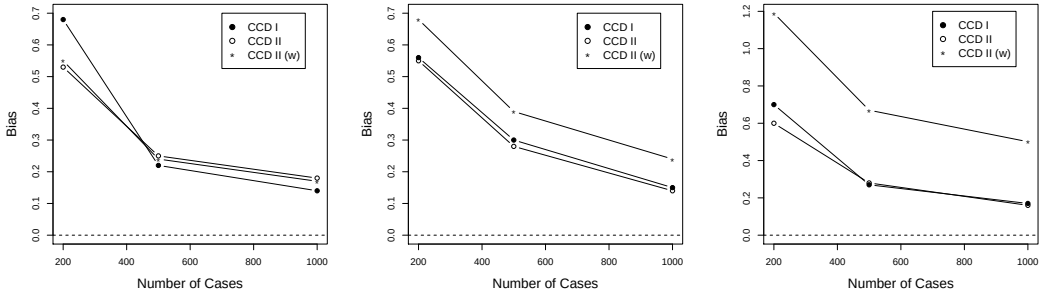
Simulation 1 also demonstrates the use of weights q_0 and $(1-q_0)\frac{1}{J}$ with matched data, for situations where $\bar{q}_0(M)$ is unknown for Case-Control Design II. This weighting scheme provided a reasonable approximation, yielding larger standard errors, but similar levels of bias for covariates with a weak association with Y. As association with Y increased, the estimate of the odds ratio became

Table 1: **Simulation 1 – Efficiency.** II MSE is Mean Squared Error for Case-Control Design II with weights $(1-q_0)\frac{1}{J}$ for CCW T-MLE, II RE is relative efficiency of Case-Control Design II CCW T-MLE with $\bar{q}_0(M)$ weights, I RE is relative efficiency of Case-Control Design I CCW T-MLE, all REs are in comparison to II MSE, and nC is Number of Cases.

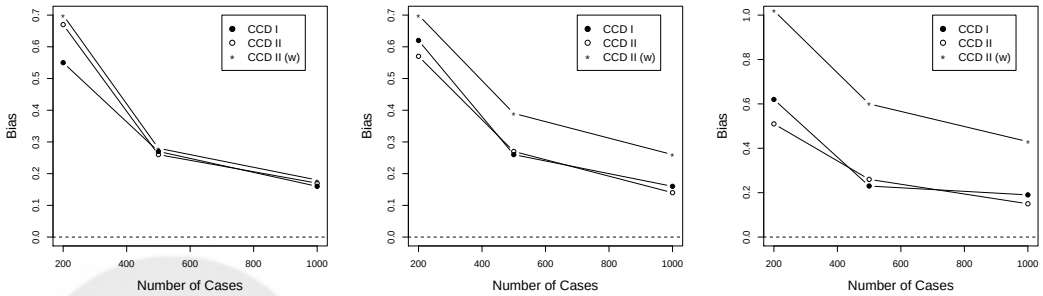
		1:1 Matching			**1:2 Matching**		
	nC	200	500	1000	200	500	1000
W_1	II MSE	2.83	0.83	0.33	1.05	0.35	0.16
	II RE	1.06	1.08	1.10	1.07	1.10	1.13
	I RE	1.15	1.14	1.13	1.04	1.06	1.12
W_2	II MSE	3.02	0.77	0.38	1.22	0.45	0.18
	II RE	1.15	1.10	1.15	1.14	1.13	1.21
	I RE	1.16	1.03	1.34	1.14	1.38	1.33
W_3	II MSE	4.67	1.40	0.60	2.07	0.71	0.41
	II RE	2.40	2.38	2.56	2.22	2.48	3.09
	I RE	1.91	1.85	2.07	2.01	2.17	3.21
W_4	II MSE	2.27	0.65	0.31	1.06	0.33	0.14
	II RE	1.03	1.02	1.02	1.01	1.02	1.01
	I RE	0.80	1.08	1.13	1.01	0.97	0.94
W_5	II MSE	2.60	0.75	0.33	1.20	0.37	0.18
	II RE	1.24	1.23	1.18	1.23	1.23	1.26
	I RE	1.01	0.99	1.11	1.11	1.04	1.31
W_6	II MSE	5.25	1.44	0.64	2.17	0.70	0.38
	II RE	2.30	2.37	2.68	2.37	2.56	3.23
	I RE	1.71	2.27	2.10	2.23	2.22	2.74
W_7	II MSE	2.63	0.70	0.31	1.10	0.33	0.16
	II RE	1.03	1.01	1.02	1.02	1.02	1.02
	I RE	1.15	0.97	1.05	1.00	1.03	1.27
W_8	II MSE	2.40	0.79	0.31	1.07	0.35	0.17
	II RE	1.20	1.30	1.43	1.25	1.41	1.54
	I RE	0.93	1.14	1.08	1.11	1.11	1.30
W_9	II MSE	4.35	1.37	0.58	1.63	0.58	0.33
	II RE	2.46	2.35	2.39	2.30	2.39	2.70
	I RE	1.76	2.13	1.90	1.45	1.83	2.49



(a) W_1 (b) W_2 (c) W_3

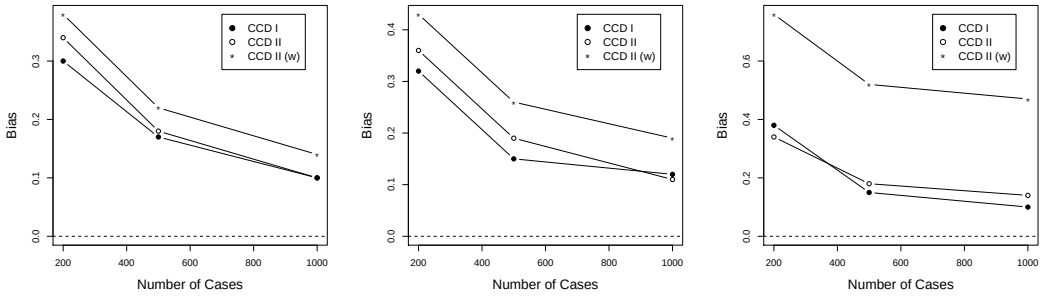


(d) W_4 (e) W_5 (f) W_6

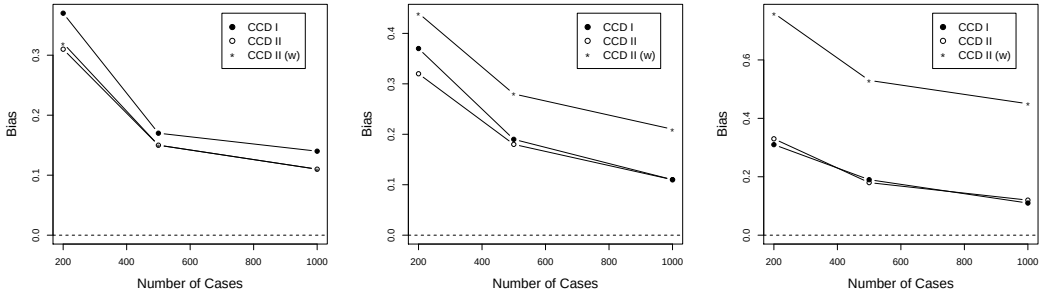


(g) W_7 (h) W_8 (i) W_9

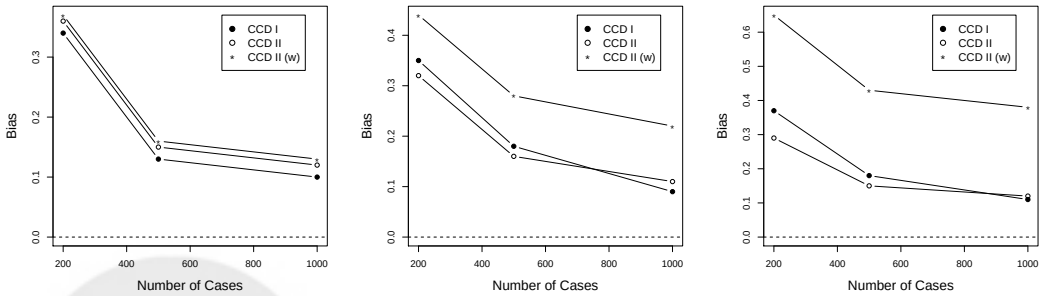
Figure 3: **Simulation 1 – Bias for 1:1 Matching.** CCD I is CCW T-MLE for Case-Control Design I, CCD II is CCW T-MLE for Case-Control Design II with $\bar{q}_0(M)$ weighting, and CCD II (w) is CCW T-MLE for Case-Control Design II with $(1 - q_0)$ weighting.



(a) W_1 (b) W_2 (c) W_3



(d) W_4 (e) W_5 (f) W_6



(g) W_7 (h) W_8 (i) W_9

Figure 4: **Simulation 1 – Bias for 1:2 Matching.** CCD I is CCW T-MLE for Case-Control Design I, CCD II is CCW T-MLE for Case-Control Design II with $\bar{q}_0(M)$ weighting, and CCD II (w) is CCW T-MLE for Case-Control Design II with $(1 - q_0)$ weighting.

more biased. Mean squared errors and relative efficiencies for the odds ratio can be seen in Table 1. Bias results can be seen in Figures 3 and 4.

5.2 Simulation 2

Our second simulation study was designed to address less ideal, and perhaps more common, situations where control information is discarded. Controls were sampled from the population of controls in Simulation 1 until a match on covariate W_i was found for each case. Non-matches were returned to the population of controls. The number of total controls sampled to find sufficient matches was recorded for each simulation. This was the number of randomly sampled controls that was used for the corresponding Case-Control Design I simulation. The mean number of controls sampled to achieve 1:1 and 1:2 matching at each sample size is noted in Table 2 as nCo. For example, in order to obtain 200 controls matched on covariate W_1 in a 1:1 design, an average of 404 controls had to be sampled from the population. Thus, an average of 404 controls were used in the corresponding Case-Control Design I.

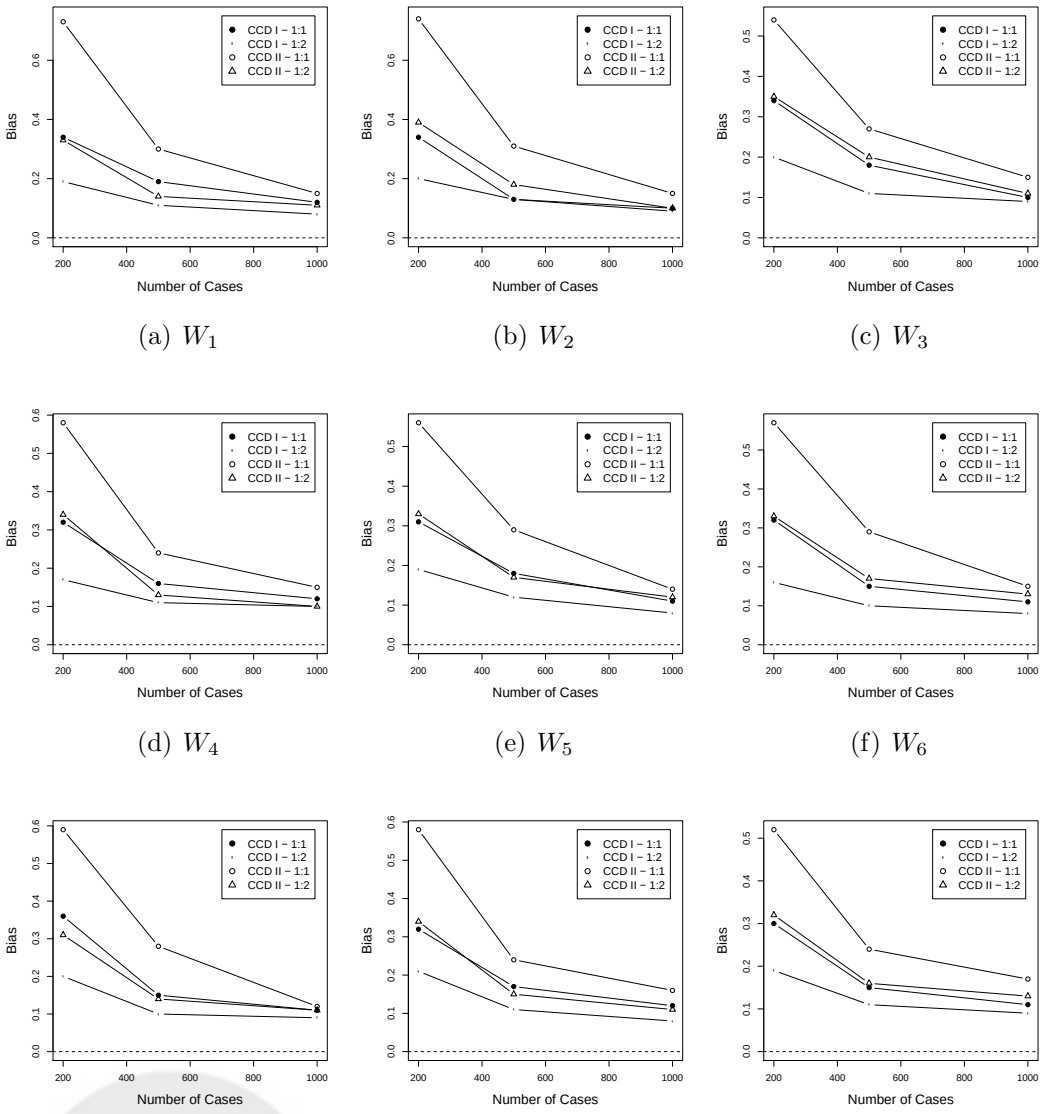
Case-control weighted targeted maximum likelihood estimation was performed for Case-Control Designs I and II. Case-Control Design I outperformed Case-Control Design II with respect to efficiency and bias for all sample sizes and both 1:1 and 1:2 matching. This was not surprising given the mean number of controls in each of the control samples for Case-Control Design I (on average, about two times the number of controls in each control sample for Case-Control Design II). Additionally, as association between W_i and Y increased, there was a trend that the number of controls necessary for complete matching also increased. A similar trend between A and W_i was not apparent. When returning to the bias results, one can see that they do not vary greatly with association between W_i and A or Y. Mean squared errors and relative efficiencies for the odds ratio can be seen in Table 2. Bias results are displayed in Figure 5.

6 Discussion

The main benefit of a matched case-control study design is a potential increase in efficiency. However, an increase in efficiency is not automatic. If one decides to implement a matched case-control study design, matching variable selection is crucial. Numerous publications in our literature review indicated that matching on non-confounding variables is not beneficial, including Kupper et al. (1981): *"The futility of matching in [non-confounding situations]*

Table 2: **Simulation 2 – Efficiency.** II MSE is Mean Squared Error for Case-Control Design II CCW T-MLE, I RE is Relative Efficiency of Case Control Design I CCW T-MLE Compared to Case-Control Design II MSE, nC is Number of Cases, nCo is Mean Number of Controls for Case-Control Design I.

		1:1 Matching			**1:2 Matching**		
	nC	200	500	1000	200	500	1000
W_1	nCo	404	1006	2010	804	2011	4026
	II MSE	2.90	0.76	0.28	1.00	0.27	0.14
	I RE	2.89	2.24	2.14	2.12	1.70	2.16
W_2	nCo	404	1009	2016	808	2016	4031
	II MSE	2.91	0.77	0.30	1.15	0.36	0.16
	I RE	2.91	2.72	2.13	2.32	2.21	2.49
W_3	nCo	406	1016	2033	812	2034	4065
	II MSE	1.99	0.48	0.22	0.84	0.28	0.11
	I RE	1.82	1.43	1.65	1.81	1.78	1.85
W_4	nCo	403	1006	2010	806	2012	4023
	II MSE	2.47	0.67	0.29	1.09	0.28	0.13
	I RE	2.38	2.09	2.20	2.29	1.91	2.03
W_5	nCo	406	1010	2019	810	2019	4040
	II MSE	2.41	0.63	0.25	0.92	0.29	0.12
	I RE	2.24	2.00	1.92	1.95	1.89	2.10
W_6	nCo	411	1025	2046	819	2045	4094
	II MSE	2.08	0.64	0.23	0.88	0.27	0.13
	I RE	2.13	1.99	1.69	1.92	1.70	2.23
W_7	nCo	402	1001	2000	801	1999	4000
	II MSE	2.71	0.72	0.30	1.09	0.34	0.15
	I RE	2.54	2.42	2.18	2.19	2.25	2.18
W_8	nCo	407	1014	2028	811	2027	4055
	II MSE	2.28	0.56	0.23	0.97	0.25	0.11
	I RE	2.35	1.76	1.71	1.99	1.59	1.68
W_9	nCo	413	1030	2059	824	2061	4121
	II MSE	1.97	0.54	0.22	0.80	0.26	0.12
	I RE	1.91	1.77	1.69	1.62	1.69	1.84



(a) W_1 (b) W_2 (c) W_3

(d) W_4 (e) W_5 (f) W_6

(g) W_7 (h) W_8 (i) W_9

Figure 5: **Simulation 2 – Bias.** CCD I is CCW T-MLE for Case-Control Design I and CCD II is CCW T-MLE for Case-Control Design II.

is clear...matching on [the variable] will have absolutely no effect on the distribution of the exposure variable in the diseased and nondiseased groups." Therefore, increases in efficiency with a matched design depend heavily on the selection of a confounding variable as a matching variable. In practice, it may be difficult to ascertain the strength of the association between the matching variable, the exposure of interest, and the outcome. Our simulations for causal effect estimation confirmed the consensus in the existing literature: that in situations where the distribution of the matching covariate is drastically different between the case and control populations, matching may provide an increase in efficiency. Our simulations indicated that $P_0^*(Y = 1 \mid W_i = 1, Z, A)$, for matching variable W_i and covariate vector Z, may need to be very small for an increase in efficiency using a matched design. These results were true, however, only for our simulations where *no control subjects were discarded*; it is very common for matched study designs to discard controls (Freedman, 1950; Cochran and Chambers, 1965; Billewicz, 1965; McKinlay, 1977).

This paper focused on the issue of individual matching in case-control studies where the researcher is interested in estimating the marginal causal effect and certain prevalence probabilities are known. Thus, we compared the use of case-control weighted targeted maximum likelihood estimation in matched and unmatched designs. We showed that in practical situations (e.g. when controls are discarded), an unmatched design is likely to be a more efficient, less biased study design choice. Since we also have a nonparametric double robust locally efficient procedure for the estimation of causal parameters in unmatched case-control study designs using q_0, it may be preferred to causal parameter estimation in matched designs. Furthermore, when q_0 is estimated, van der Laan (2008) demonstrated that one can incorporate the uncertainty surrounding the estimate of q_0 into the standard error of the parameter of interest. However, if controls will not be discarded, there is a priori information about the matching variable(s), or the circumstances only allow for a matched design, our double robust locally efficient procedure for the estimation of causal parameters in matched case-control study designs can then be used, as demonstrated in this paper. This design relies on the additional knowledge of $\bar{q}_0(M)$. Our simulations also indicated that when $\bar{q}_0(M)$ is unknown, $1 - q_0$ may provide a reasonable approximation, although this should be examined further.

References

O. Bembom, M.L. Peterson, S-Y Rhee, W.J. Fessel, S.E. Sinisi, R.W. Shafer, and M.J. van der Laan. Biomarker discovery using targeted maximum likelihood estimation: Application to the treatment of antiretroviral resistant hiv infection. *Technical Report 221, Division of Biostatistics, University of California, Berkeley*, 2007.

J. Benichou and S. Wacholder. A comparison of three approaches to estimate exposure-specific incidence rates from population-based case-control data. *Statistics in Medicine*, 13:651–661, 1994.

W.Z. Billewicz. The efficiency of matched samples: An empirical investigation. *Biometrics*, 21(3):623–644, 1965.

N.E. Breslow and N.E. Day. *Statistical Methods in Cancer Research: Volume 1 – The analysis of case-control studies*. International Agency for Research on Cancer, Lyon, 1980.

N.E. Breslow, N.E. Day, K.T. Halvorsen, R.L. Prentice, and C. Sabal. Estimation of multiple relative risk functions in matched case-control studies. *Am J Epid*, 108(4):299–307, 1978.

W.G. Cochran. Matching in analytical studies. *American Journal of Public Health*, 43:684–691, 1953.

W.G. Cochran and S.P. Chambers. The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A (General)*, 128(2):234–266, 1965.

D. Collett. *Modeling Binary Data*. Chapman and Hall, London, 1991.

M.C. Costanza. Matching. *Preventive Medicine*, 24:425–433, 1995.

R. Freedman. Incomplete matching in ex post facto studies. *The American Journal of Sociology*, 55(5):485–487, 1950.

O. Gefeller, A. Pfahlberg, H. Brenner, and J. Windeler. An empirical investigation on matching in published case-control studies. *European Journal of Epidemiology*, 14:321–325, 1998.

S. Greenland. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *Am J Epidemiol*, 160(4):301–305, 2004.

S. Greenland. Multivariate estimation of exposure-specific incidence from case-control studies. *J Chron Dis*, 34:445–453, 1981.

T.R. Holford, C. White, and J.L. Kelsey. Multivariate analysis for matched case-control studies. *Am J Epid*, 107(3):245–255, 1978.

P.W. Holland and D.B. Rubin. Causal inference in retrospective studies. In D.B. Rubin, editor, *Matched Sampling for Causal Effects.* Cambridge University Press, Cambridge, MA, 1988.

L.L. Kupper, J.M. Karon, D.G. Kleinbaum, H. Morgenstern, and D.K. Lewis. Matching in epidemiologic studies: Validity and efficiency considerations. *Biometrics*, 37:271–291, 1981.

S.M. McKinlay. Pair-matching – a reappraisal of a popular technique. *Biometrics*, 33(4):725–735, 1977.

O.S. Miettinen. Estimation of relative risk from individually matched series. *Biometrics*, 26:75–86, 1970.

K.L. Moore and M.J. van der Laan. Covariate adjustment in randomized trials with binary outcomes. *Technical Report 215, Division of Biostatistics, University of California, Berkeley*, 2007.

M. Rahman. Analysis of matched case-control data: Author reply. *J of Clin Epidemiol*, 56(8):814, 2003.

S. Rose and M.J. van der Laan. Simple optimal weighting of cases and controls in case-control studies. *The International Journal of Biostatistics*, 4(1): Article 19, 2008.

K. Rothman and S. Greenland. *Modern Epidemiology.* Lippincott, Williams and Wilkins, Philadelphia, PA, 2nd edition, 1998.

D.B. Rubin. *Matched Sampling for Causal Effects.* Cambridge University Press, Cambridge, MA, 2006.

J.J. Schlesselman. *Case-Control Studies: Design, Conduct, Analysis.* Oxford University Press, Oxford, 1982.

D.G. Seigel and S.W. Greenhouse. Validity of estimating relative risk in case-control studies. *J Chron Dis*, 26:219–225, 1973.

M.J. van der Laan. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1):Article 2, 2006.

M.J. van der Laan. Estimation based on case-control designs with known prevalence probability. *The International Journal of Biostatistics*, 4(1): Article 17, 2008.

M.J. van der Laan and J.M. Robins. *Unified methods for censored longitudinal data and causailty*. Springer, New York, 2002.

M.J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):Article 11, 2006.

J.P. Vandenbroucke, E. von Elm, D.G. Altman, P.C. Gotzsche, C.D. Mulrow, S.J. Pocock, C. Poole, J.J. Schlesselman, and M. Egger for the STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and elaboration. *PLoS Medicine*, 4(10):1628–1654, 2007.

8.4 *Causal Inference for Nested Case-Control Studies using Targeted Maximum Likelihood Estimation*

The following article appears as it was originally published on the University of California, Berkeley Division of Biostatistics Working Paper Series website in 2009, http://www.bepress.com/ucbbiostat/paper253/.

Causal Inference for Nested Case-Control Studies using Targeted Maximum Likelihood Estimation

Sherri Rose and Mark J. van der Laan

Abstract

A nested case-control study is conducted within a well-defined cohort arising out of a population of interest. This design is often used in epidemiology to reduce the costs associated with collecting data on the full cohort; however, the case control sample within the cohort is a biased sample. Methods for analyzing case-control studies have largely focused on logistic regression models that provide conditional and not marginal causal estimates of the odds ratio. We previously developed a Case-Control Weighted Targeted Maximum Likelihood Estimation (TMLE) procedure for case-control study designs, which relies on the prevalence probability q_0. We propose the use of Case-Control Weighted TMLE in nested case-control samples, with either known q_0 or q_0 estimated from the full cohort. We show that this procedure is efficient for a reduced data structure, the data structure where covariate information is not collected or available on non-case-control subjects, and recognize that it is not fully efficient for the full data. However, in many common scenarios, the full data is not available, thus our procedure is maximally efficient for the data given. For statistical inference, we view the nested case-control sample as a missing data problem (Robins et al., 1994). Case-Control Weighted TMLE on the reduced data structure is illustrated in simulations for cohorts with and without right censoring and also effect modification in randomized controlled trials.

Keywords: nested case control sampling, causal effect, counterfactual, double robust estimation, estimating function, inverse probability of treatment weighting, locally efficient estimation, marginal structural models, targeted maximum likelihood estimation, treatment effect, variable importance measures

1 Introduction

Nested case-control studies are conducted within a well-defined cohort arising out of a population of interest. Typically, all of the subjects that develop disease in the cohort (i.e., the cases) are selected along with a random sampling of non-diseased subjects. Controls may be selected at the time each case becomes a case from the population without an event at that time but at risk for the event or at the end of the study. These two groups of subjects then comprise the nested case-control sample, where it is common for additional information to be collected, such as the exposure of interest (Mantel, 1973; Kupper et al., 1975; Liddell et al., 1977; Breslow et al., 1983; Rothman and Greenland, 1998). This design is increasingly used in public health, medicine, and genomics to study relationships between exposures and disease in large observational cohorts and effect modification in randomized controlled trials (Rothman and Greenland, 1998; Essebag et al., 2003, 2005). Nested designs may reduce the costs associated with collecting data on the full cohort with only a nominal loss in efficiency (Ernster, 1994; Rothman and Greenland, 1998; Hak et al., 2004; Vittinghoff and Bauer, 2006).

However, whether nested within a large observational cohort or a randomized controlled trial, the case-control study nested within the full cohort is biased since the proportion of cases in the sample is not the same as the population of interest. Methods for analyzing case-control studies have largely focused on logistic regression models (Breslow and Cain, 1988). These models provide conditional and not marginal (causal) estimates of the odds ratio. We have developed a Case-Control Weighted Targeted Maximum Likelihood Estimation (TMLE) procedure for case-control samples, which relies on the prevalence probability $q_0 \equiv P_0^*(Y = 1)$. TMLE is a general procedure for estimation, and can be used for any full data model and parameter of interest. It is a two-step method where one first obtains an estimate of the data-generating distribution and then in second stage updates the initial fit in a bias-reduction step targeted towards the parameter of interest, instead of the overall density. For case-control data, we simply employ the use of case-control weights in Case-Control Weighted TMLE. We propose the extension of Case-Control Weighted TMLE in nested case-control samples, with either known q_0 or q_0 estimated from the full cohort. We show that this procedure is efficient for a reduced data structure, the data structure where covariate information is not collected or available on non-case-control subjects, and recognize that it is not fully efficient for the full data. However, in many common scenarios, the full data is not available, thus our procedure is maximally efficient for the data given. For statistical inference, we view the nested case-control sample as a

missing data problem Robins et al. (1994). We are able to estimate a variety of parameters with Case-Control Weighted TMLE, including the marginal exposure effect adjusted for confounders. These parameters can be viewed as the analogues of causal inference parameters, but for observational data. We refer to these parameters as variable importance parameters if we are not willing to make causal assumptions. We illustrate Case-Control Weighted TMLE on the reduced data structure in simulations for cohorts with and without right censoring and also effect modification in randomized controlled trials.

2 Background

2.1 Literature and Existing Methodology

Nested case-control studies were introduced in Mantel (1973) and further discussed and developed in Kupper et al. (1975), Liddell et al. (1977), Thomas (1977), and Breslow et al. (1983). Advantages include reduction in costs associated with collecting data on the entire cohort, minimal losses in efficiency, and having the cases and controls come from the same population (Ernster, 1994; Rothman and Greenland, 1998; Essebag et al., 2003; Hak et al., 2004; Vittinghoff and Bauer, 2006). The latter is frequently not the case in independent case-control study designs. Nested case-control designs have also been shown to have similar estimates for parameters such as the standardized morbidity ratio when compared to an analysis of the full cohort (Liddell et al., 1977; Breslow et al., 1983; Lubin, 1986).

Much of the literature for analysis of nested case-control studies focuses on logistic regression models. The use of conditional logistic regression, treating the nested case-control study as a sample matched on time, is frequently discussed (Breslow and Cain, 1988; Flanders and Greenland, 1991; Ernster, 1994; Barlow et al., 1999; Szklo and Nieto, 1999). Samuelsen (1997) constructs pseudolikelihoods for nested case-control study designs using the conditional probability that a subject will be selected as a control to build a general parametric regression estimator and a semiparametric proportional-hazards estimator. Proportional hazards models have also been discussed elsewhere (e.g., Lubin, 1986). An important reference for our methodology is Robins et al. (1994). Their paper includes a discussion of a missingness framework for the estimation of inverse probability of treatment weighted (IPTW) marginal causal parameters for nested case-control study designs. We also refer to van der Laan and Robins (2003) which handles double robust estimation for missing data structures.

Beyond the types of parameters being estimated, the literature on the analysis of nested case-control study designs could further be divided loosely into three groups. One group analyzes the nested case-control sample as a case-control sample, ignoring the first stage of sampling the cohort, for example Barlow et al. (1999). The second group analyzes the nested case-control sample as a missing data structure, such as Robins et al. (1994). The third group straddles both of these groups, for example Breslow and Cain (1988). Our methodology falls within this third group. We estimate our parameter with information from only the case-control sample, but our inference respects the missing data structure. Our variance estimates incorporate both the variability due to sampling the cohort from the population of interest and the variability arising from drawing the case-control sample from the cohort.

An additional division in the literature could be drawn based on methods that rely on knowledge of the prevalence probability $q_0 \equiv P_0^*(Y = 1)$. For example, the methodology of Robins et al. (1994) requires only that q_0 be small. Our proposed methodology uses knowledge of q_0, or a reasonable estimate of q_0 approximated within the full cohort. The use of q_0 to eliminate the bias of case-control sampling designs has previously been discussed as update to a logistic regression model with the intercept $\log q_0/(1-q_0)$ (Anderson, 1972; Prentice and Breslow, 1978; Greenland, 1981; Morise et al., 1996; Wacholder, 1996). Adding the intercept $\log q_0/(1-q_0)$ yields the true logistic regression function $P_0^*(Y = 1 \mid A, W)$ (Anderson, 1972; Prentice and Pyke, 1979). A discussion of this updated logistic regression and its sensitivity to model misspecification can be found in Rose and van der Laan (2008). Similarly, there is a wealth of literature which discusses estimation in nested case-control studies with known sampling probabilities from the cohort, such as Borgan and Langholz (1993).

2.2 Case-Control Weighted TMLE

TMLE is a general methodology introduced in van der Laan and Rubin (2006). It is an efficient and double robust procedure that can estimate a variety of parameters of interest. We propose the use of Case-Control Weighted TMLE, which is simply a TMLE procedure that relies on the prevalence probability for case-control weights, in the case-control observations nested within a cohort. We will view the nested case-control sample within the cohort as a biased case-control sample in order to estimate our parameter of interest. Thus, here we discuss the general methodology for Case-Control Weighted TMLE before describing its application for use in nested case-control studies.

Case-Control Weighted TMLE, discussed in van der Laan (2008), maintains the locally efficient double robustness properties of estimating function based

methodology, and unifies maximum likelihood estimation (MLE) with estimating function methodology into a method improving on both. The case-control weighting framework maps estimation methods designed for non-case-control sampling into methods for case-control sampling. Case-control weighting allows us to provide TMLE methodology, which targets the parameter of interest, for biased case-control sampling in the form of Case-Control Weighted TMLE. Our procedure is a general methodology for the estimation of a parameter of a probability distribution, such as marginal causal effects and variable importance measures. The methodology relies on knowledge of the true prevalence probability $P_0^*(Y = 1) \equiv q_0$, or a reasonable approximation, to eliminate the bias of the case-control sampling design.

Let us define $O^* \sim P_0^*$ as the experimental unit and corresponding distribution P_0^* of interest. To generalize, our case-control weighting maps a function of O^* into a function of the case-control data structure O, while preserving the expectation of the function. For example, the experimental unit of interest may be defined as $O^* = (W, A, Y) \sim P_0^*$, which consists of baseline covariates W, an exposure variable A, and a binary outcome Y. Then, in an independent case-control study design sampling can be described as first sampling (W_1, A_1) from the conditional distribution of (W, A), given $Y = 1$ for a case and then sampling J controls (W_0^j, A_0^j) from (W, A), given $Y = 0, j = 1, \ldots, J$. The observed data structure O is then defined by:

$$O = ((W_1, A_1), (W_0^j, A_0^j : j = 1, \ldots, J)) \sim P_0, \text{ with}$$

$$(W_1, A_1) \sim (W, A \mid Y = 1)$$
$$(W_0^j, A_0^j) \sim (W, A \mid Y = 0),$$

where the cluster containing one case and J controls is considered the experimental unit, and the marginal distribution of this cluster is specified by P_0^*. A case-control dataset of this design then consists of n i.i.d. observations $O_1, \ldots, O_n$ with sampling distribution P_0 as described above. The model $\mathcal{M}^*$, where q_0 may or may not be known, implies models for the marginal distribution of cases (W_1, A_1) and controls $(W_2^j, A_2^j), j = 1, \ldots, J$. Of note, if the independent case-control sampling design is conducted simply as sampling nC cases from the conditional distribution of (W, A), given $Y = 1$, and sampling nCo controls from (W, A), given $Y = 0$, the value of J used to weight each control is then nCo/nC.

Let $O^* \to D^*(O^*)$ represent an estimating function or loss function for O^* that can be used to estimate the parameter of interest of P_0^* based on an i.i.d. sample of O^*. We are concerned with mapping this function D^* into a function

for this same parameter of interest, but now based on sampling O (a biased sample for O^*). We define the case-control weighted version:

$$D_{q_0}(O) \equiv q_0 D^*(W_1, A_1, 1) + \frac{1}{J}\sum_{j=1}^{J}(1-q_0)D^*(W_2^j, A_2^j, 0),$$

which is now a function of the observed experimental unit O. Additionally, we define the expectation operator $P_{0,q_0}D^* = P_0 D_{q_0}$, which takes the expectation of the case-control weighted function $D_{q_0}(O)$ with respect to P_0. Similarly, we define the empirical expectation $P_{n,q_0}D^* = P_n D_{q_0}$ as the empirical mean of the case-control weighted D_{q_0}, where P_n is the empirical distribution of $O_1, \ldots, O_n$. Now, we can let $D^*(O^*)$ be a function so that $P_0^* D^* \equiv E_{P_0^*} D^*(O^*) = 0$. Then $P_0 D_{q_0} = 0$, and

$$D_{q_0}(O) \equiv q_0 D^*(W_1, A_1, 1) + (1-q_0)\frac{1}{J}\sum_{j=1}^{J} D^*(W_2^j, A_2^j, 0).$$

In more generality, for any function D^* and corresponding case-control weighted function D_{q_0}, we have

$$P_0 D_{q_0} = P_0^* D^*.$$

Given a model $\mathcal{M}^*$ for p_0^*, we can estimate P_0^* with a case-control weighted maximum likelihood estimator:

$$p_n^* = \arg\max_{p^* \in \mathcal{M}^*} \sum_{i=1}^{n} L(O_i, p^*),$$

where $L(O_i, p^*)$ is the case-control weighted log likelihood loss function for the density p_0^* of O^* under sampling of $O \sim P_0$:

$$L(O_i, p^*) = q_0 \log p^*(W_1, A_1, 1) + (1-q_0)\frac{1}{J}\sum_{j=1}^{J} \log p^*(W_2^j, A_2^j, 0).$$

Now, let $D^*(P_0^*)$ be the efficient influence curve of the parameter $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$. We consider an initial estimator P_n^{*0} of P_0^* based on $O_1, \ldots, O_n$ such as a case-control weighted maximum likelihood estimator according to a working model within $\mathcal{M}^*$. Let $\{P_n^*(\epsilon) : \epsilon\}$ be a submodel of $\mathcal{M}^*$ with parameter ϵ satisfying that the linear span of its score at $\epsilon = 0$ includes $D^*(P_n^{*0})$. Then we let ϵ_n^1 be the case-control weighted maximum likelihood estimator of ϵ:

$$\epsilon_n^1 = \arg\max P_{n,q_0} \log p_n^{*0}(\epsilon).$$

From this we obtain an update $P_n^{*1} = P_n^{*0}(\epsilon_n^1)$ of the initial estimator P_n^{*0}. This updating process is iterated until step k at which $\epsilon_n^k \approx 0$. The final update is denoted P_n^*. By the score condition, this final estimator solves the case-control weighted efficient influence curve:

$$0 = P_{n,q_0} D^*(P_n^*) = P_n D_{q_0}(P_n^*)$$

up to numerical precision (van der Laan and Rubin, 2006). We refer to $\psi_n = \Psi^*(P_n^*)$ as the case-control weighted targeted maximum likelihood estimator of ψ_0.

The theoretical development of Case-Control Weighted TMLE can be found in van der Laan (2008). In Rose and van der Laan (2008), we implemented Case-Control Weighted TMLE and presented a comparison of the procedure to an existing method for estimation of the causal parameters in case-control studies, the approximately correct IPTW of Robins (1999). We demonstrated that Case-Control Weighted TMLE outperforms the IPTW method for estimation of the marginal causal odds ratio in many practical situations.

3 Methodology for Nested Designs

Our goal is to apply Case-Control Weighted TMLE methodology to nested case-control designs. First, it is important to understand the statistical framework for the design. Nested case-control study designs have a missing data structure, as presented by Robins et al. (1994), and which we will discuss here. We will use a reduced data structure to estimate the parameter of interest with our proposed case-control weighted targeted maximum likelihood estimator. This estimator solves the efficient influence curve equation for the reduced data structure.

3.1 The Data Structure

Let O^* be a full data structure of the experimental unit O^* represents the data that ideally would be observed in order to answer the research question of interest. In most studies, however, one or more components of the full data are subject to one or more types of missingness, and only $O = \Phi(O^*, \delta)$ can be observed, where Φ is a known many-to-one mapping and δ denotes a missingness variable. Here, O^* represents data from the full cohort data and the missingness variable indicates membership in the nested case-control sample.

Suppose the full data structure is $O^* = (W, A, Y)$ with Y being a binary outcome of interest, A a binary exposure, and W a vector of covariates. Let us also suppose that the observed data structure for the nested case-control study is $O = (\delta, \delta O_1^*, O_2^*)$, where $O^* = (O_1^*, O_2^*)$. Particular examples are that $O_1^* = A$ and $O_2^* = (W, Y)$, or $O_1^* = (A, W)$, and $O_2^* = Y$. It is assumed that O_2^* always includes Y. The observations with $\delta = 1$ are the observations in the nested case-control sample within the cohort and have additional variables O_1^* measured. If $O_2^* = Y$, the missing data structure essentially ignores the non-case-control observations, except for the purpose of estimating $q_0 \equiv P_0^*(Y = 1)$. Covariate and exposure information is not available or is not measured. This case is particularly interesting since we can show that the case-control weighted targeted maximum likelihood estimator using only the case-control observations and the empirical estimate of q_0 obtained from the full cohort is a targeted maximum likelihood estimator for this particular missing data structure $(\delta, \delta(W, A), Y)$. If covariate information is measured and available for non-case-control subjects, this missing data structure ignores the information and therefore our estimator is not fully efficient.

We assume the coarsening at random (CAR) assumption: $\Pi(O^*) \equiv P_0^*(\delta = 1 \mid O^*) = P_0^*(\delta = 1 \mid O_2^*)$, and a special case is that $P_0^*(\delta = 1 \mid O_2^*) = P_0^*(\delta = 1 \mid Y)$ with $P_0^*(\delta = 1 \mid Y = 1) = 1$ and $P_0^*(\delta = 1 \mid Y = 0) = p$, where p is estimated empirically from the data. In this case the selection for the case-control sample is based upon the outcome Y. One might wish to choose p so that a single case $(Y = 1, \delta = 1)$ corresponds with J-controls $(Y = 0, \delta = 1)$, on average. If $q_0 = P_0^*(Y = 1)$, then $Jq_0 P_0^*(\delta = 1 \mid Y = 1) = (1 - q_0) P_0^*(\delta = 1 \mid Y = 0)$, which results in $p = \frac{Jq_0}{1-q_0}$.

3.2 Parameter of Interest

The statistical problem is then to estimate the parameter $\psi_0 = \Psi^*(P_0^*)$ of the population distribution $P_0^* \in \mathcal{M}^*$ of (W, A, Y), known to be an element of some specified model $\mathcal{M}^*$, based on the nested case-control data set $O_1, \ldots, O_n \sim P_0$. $O^* \sim P_0^*$, the experimental unit of interest, is not the observed experimental unit, due to the missing data structure. P_0^* now represents the full data distribution and P_0 is the distribution of the missing data structure with observed experimental unit $O = (\delta, \delta O_1^*, O_2^*) \sim P_0$. We focus on the case $O_2^* = Y$, where covariate information on non-case-control subjects is unavailable or ignored, and view this missing data structure as a biased case-control sampling design in order to estimate our parameter of interest. An example of a parameter of interest is the marginal exposure effect on the additive scale, which can also

be viewed as the causal risk difference:

$$\begin{aligned}\psi_{0,RD} &\equiv E_0^*\{E_0^*(Y \mid A=1,W) - E_0^*(Y \mid A=0,W)\} \\ &= E_0^*(Y_1) - E_0^*(Y_0) = P_0^*(Y_1=1) - P_0^*(Y_0=1).\end{aligned}$$

This definition requires the specification of the counterfactual outcomes Y_0 and Y_1 for binary A and $(W, A, Y = Y_A)$ as a time-ordered missing data structure on (W, Y_0, Y_1). For a causal interpretation, one must also make the randomization assumption: $\{A \perp Y_0, Y_1 \mid W\}$, meaning there are no unmeasured confounders. This parameter can also be viewed as a W-adjusted variable importance parameter, as previously mentioned, without the need to make causal assumptions. See van der Laan (2006) for this framework. We make use of the shorthand $Q_0^* = P_0^*(Y \mid A, W)$ and $g_0^* = P_0^*(A \mid W)$, the latter often referred to as the "treatment mechanism" but as the "exposure mechanism" in case-control studies.

3.3 The Estimator

TMLE is a general procedure for estimation, and can be used for any full data model and parameter of interest. It is a two-step method where one first obtains an estimate of the data-generating distribution and then in second stage updates the initial fit in a bias-reduction step targeted towards the parameter of interest, instead of the overall density. For case-control data we then simply add case-control weighting, using the prevalence probability. Here we will use Case-Control Weighted TMLE applied to nested case-control data using an estimate of q_0 from the full cohort. Again we focus on the case where O_2^* in the experimental unit $O = (\delta, \delta O_1^*, O_2^*) \sim P_0$ is equal to Y. We can show that the case-control weighted targeted maximum likelihood estimator using only the case-control observations and the empirical estimate of q_0 obtained from the full cohort is a targeted maximum likelihood estimator for this particular missing data structure $(\delta, \delta(W, A), Y)$. In this special case, the $D^*(Q, g, \Pi)$ we solve is the efficient influence curve (see Section 3.4). In other cases, for example when $O_2^* = (W, Y)$, we follow the same template for targeted maximum likelihood, where the case-control weighted log-likelihood is the criterion for fit.

Let us say we are still interested in the risk difference. We also let Q_n^0 be an initial estimator of $Q_0^* = P_0^*(Y \mid A, W)$, say the case-control weighted maximum likelihood estimator, or equivalently, the inverse probability of censoring weighted (IPCW) logistic regression estimator. In the IPCW estimator, the weights are $\frac{\delta}{\Pi}$. Now, we construct the ϵ-extension $\text{logit}Q_n^0(\epsilon) =$

$\text{logit}Q_n^0 + \epsilon h(g)(A,W)$, where $h(g)(A,W) \equiv \frac{A}{g_0^*(1|W)} - \frac{1-A}{g_0^*(0|W)}$, and we estimate ϵ with IPCW MLE. Alternatively, one puts the inverse probability of censoring weights in the ϵ-covariate: $\epsilon h(g)(A,W)\frac{\delta}{\Pi}$. Let $Q_n = Q_n^0(\epsilon_n)$. We now solve $\psi_{n,RD} = P_n \frac{\delta}{\Pi}(Q_{1n} - Q_{0n})$, where $Q_{1n} = Q_n(1,W)$ and $Q_{0n} = Q_n(0,W)$. Note that this corresponds with the case-control weighted empirical mean over W. So this estimator $\psi_{n,RD}$ corresponds exactly with the case-control weighted targeted maximum likelihood estimator proposed in van der Laan (2008), Rose and van der Laan (2008), and Rose and van der Laan (2009).

3.4 The Efficient Influence Curve

In order to estimate our parameter of interest, we view the missing data structure $(\delta, \delta(W,A), Y)$, where covariate information on subjects outside the nested case-control sample is unavailable or discarded, as a case-control sample. However, inference for this parameter must respect the missing data structure in order to account for the two sources of variability in the estimator. The first source of variance arises due to drawing the cohort from the target population, and the second source of variance arises from drawing the case-control sample from the cohort. If our inference treated the sample simply as a case-control sample, we would not be incorporating the additional variance arising from sampling the cohort from the population. Thus, for inference, we use an efficient influence curve that respects the missing data structure to obtain an estimate of the variance of our estimator. However, the efficient influence curve can also be used to construct closed form locally efficient double robust estimators by using it as an estimating function. The case-control weighted targeted maximum likelihood estimator discussed in the previous section solves the efficient influence curve equation for the missing data structure $(\delta, \delta(W,A), Y)$.

Our methodology for independent case-control study designs relies on knowledge of q_0, or a reasonable approximation of q_0, for appropriate statistical inference. In nested case-control samples we can easily estimate q_0 from the full cohort data. Inference for nested case-control study designs also requires the CAR assumption: $\Pi(O^*) \equiv P_0^*(\delta = 1 \mid O^*) = P_0^*(\delta = 1 \mid O_2^*)$. Let us say that we are still interested in the risk difference, but note that the derivation of the efficient influence curve and corresponding estimators generalizes to all other parameters of the full data distribution.

The efficient influence curve in the nonparametric full data model for $O^* = (W, A, Y)$ is given by:

$$D(O^*) = h(g)(A,W)(Y - Q(A,W)) + Q(1,W) - Q(0,W) - \Psi(Q),$$

where $h(g)(A,W) \equiv \frac{A}{g(1|W)} - \frac{(1-A)}{g(0|W)}$. We will represent $D = D_1 + D_2$, where

$D_1 = h(g)(Y - Q)$. The efficient influence curve for the missing data model is obtained through the following doubly robust IPCW mapping applied to the full data efficient influence curve D (see van der Laan and Robins (2003)):

$$D^* = \frac{\delta}{\Pi}\{D - E(D \mid \delta = 1, O_2^*)\} + E(D \mid \delta = 1, O_2^*).$$

This efficient influence curve can now be used to construct closed form locally efficient double robust estimators by using it as an estimating function. One will also be able to construct corresponding targeted maximum likelihood estimators. Here, we will focus on the $O_2^* = Y$-case. We have

$$\begin{aligned} D^*(Q, g, \Pi) &= \frac{\delta}{\Pi}\{h(Y - Q) + (Q_1 - Q_0)\} \\ &\quad -\frac{\delta}{\Pi}E(h(Y - Q) + Q_1 - Q_0 \mid \delta = 1, Y) \\ &\quad +E(h(Y - Q) + Q_1 - Q_0 \mid \delta = 1, Y) - \Psi(Q), \end{aligned}$$

where we use the notation $Q_1(W) = Q_0^*(1, W)$, $Q_0(W) = Q_0^*(0, W)$, and $Q = Q_0^*(A, W)$. This efficient influence curve can be decomposed as the sum of the following two components:

$$\begin{aligned} D_1^* &= \frac{\delta}{\Pi}\left(h(Y - Q) - E(h(Y - Q) \mid \delta = 1, Y)\right) + E(h(Y - Q) \mid \delta = 1, Y) \\ D_2^* &= \frac{\delta}{\Pi}\left(Q_1 - Q_0 - E(Q_1 - Q_0 \mid \delta = 1, Y)\right) + E(Q_1 - Q_0 \mid \delta = 1, Y) - \Psi(Q). \end{aligned}$$

We claim that D_1^* is a score of $dP(Y \mid A, W)$ and D_2^* is a score of $dP(W)$ in the observed likelihood factorization of $(\delta, \delta(W, A), Y)$, where the conditional expectation contributions, given $(\delta = 1, Y)$, are coming from the $dP(Y)$-factor.

Viewing $D^* = D^*(Q^*, g^*, \Pi, \psi)$ as an estimating function in ψ, setting $P_n D^*(Q_n, g_n, \Pi, \psi_n) = 0$ for given estimators Q_n, g_n of Q_0, g_0, yields the solution for the risk difference:

$$\begin{aligned} \psi_n &= P_n \frac{\delta}{\Pi}\{h(g_n)(Y - Q_n) + (Q_{1n} - Q_{0n})\} \\ &\quad -\left(\frac{\delta}{\Pi} - 1\right)\{E_n(h(g_n)(Y - Q_n) + Q_{1n} - Q_{0n} \mid \delta = 1, Y)\}. \end{aligned}$$

It is necessary for us to estimate the nuisance parameters:

$$\begin{aligned} E_n(h(Y - Q_n) \mid \delta = 1, Y = y) &= \frac{\sum_{i=1}^n I(\delta_i = 1, Y_i = y)h(A_i, W_i)(y - Q_n(W_i, A_i))}{\sum_{i=1}^n I(\delta_i = 1, Y_i = y)} \\ E_n(Q_{1n} - Q_{0n} \mid \delta = 1, Y = y) &= \frac{\sum_{i=1}^n I(\delta_i = 1, Y_i = y)(Q_{1n} - Q_{0n})(W_i)}{\sum_{i=1}^n I(\delta_i = 1, Y_i = y)}. \end{aligned}$$

Our case-control weighted targeted maximum likelihood estimator solves the IPCW weighted efficient influence curve equation:

$$0 = P_n \frac{\delta}{\Pi} \{h(g_n)(Y - Q_n) + (Q_{1n} - Q_{0n}) - \Psi(Q_n)\}.$$

In our case-control study nested within the cohort sample, we estimate q_0 with $q_{0n} = \frac{1}{n}\sum_i I(Y_i = 1)$ and use the corresponding Π_n. Suppose we estimate $\Pi = P(\delta = 1 \mid Y = y)$ with the empirical proportion of δ among the observations with $Y_i = y$. Then:

$$0 = P_n \left(\frac{\delta}{\Pi_n} - 1 \right) \{E_n(h(g_n)(Y - Q_n) + Q_{1n} - Q_{0n} - \psi_n \mid \delta = 1, Y)\}.$$

This follows by first conditioning on $Y = y$, and then noting that $P_n(\delta/\Pi_n(y) - 1 \mid Y = y) = 0$ for each $y \in \{0, 1\}$. By estimating Π with the empirical distribution of δ, it follows that this targeted maximum likelihood estimator ψ_n solves the efficient influence curve equation:

$$0 = P_n D^*(Q_n, g_n, \Pi_n, \psi_n).$$

Thus, our case-control weighted targeted maximum likelihood estimator, using the empirical proportions from the total cohort sample for q_0 and $1 - q_0$, actually solves this efficient influence curve equation for the missing data structure $(\delta, \delta(W, A), Y)$. In particular, we can use $D^*(Q^*, g^*, \Pi, \psi)$ as the influence curve under the assumption that g_0^* is correctly estimated. This influence curve can then be used calculate standard errors of the case-control weighted targeted maximum likelihood estimator. An estimate of the asymptotic variance of $\sqrt{n}(\psi_{n,RD} - \psi_0)$ using the efficient influence curve $D^*(Q^*, g^*, \Pi, \psi)$ is given by $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \hat{D}^{*2}$. A 95% Wald-type confidence interval for a parameter estimate $\hat{\psi}$ can be constructed as: $\hat{\psi} \pm z_{0.975}\frac{\hat{\sigma}}{\sqrt{n}}$ with a p-value calculated as $2[1 - \Phi(|\frac{\hat{\psi}}{\hat{\sigma}/\sqrt{n}}|)]$. Resampling based methods can also be implemented to estimate the standard error of the estimated parameter of interest.

We conclude that our proposed case-control weighted targeted maximum likelihood estimator with the empirical q_{0n} is a targeted maximum likelihood estimator for the missing data structure $(\delta, \delta(W, A), Y)$, and is thus a locally efficient procedure for that data. If in truth, as may often be the case, the non-case-control observations have covariate data, then one can use a more efficient double robust estimator using the above efficient influence curve and estimating the nuisance parameters.

4 Right Censoring

Let us say that our full data structure (the cohort) is a censored data structure. For example, O^* might be defined as $O^* = (W, A, \tilde{T}, \Delta, Y^*)$, where:

$$
\begin{gathered}
W \text{ are covariates,} \\
A \text{ is an exposure of interest,} \\
\tilde{T} = \min(T, C), \\
T \text{ is the time to the event } Y, \\
C \text{ denotes a censoring variable,} \\
\Delta = I(\tilde{T} = T), \text{ and} \\
Y^* = (\tilde{T} \leq t, \Delta = 1).
\end{gathered}
$$

We can apply our case-control weights to any data structure, and therefore O^* can be a censored data structure and we are still able to use our methods. Thus, suppose our observed data for this full data O^* is then $O = (\delta, \delta(W, A), \tilde{T}, \Delta, Y^*)$. Again, $\delta = 1$ denotes membership in the nested case-control sample. A special feature of this right censored data structure is that the true Y is not observed or a part of the full data. Instead, as noted, we have $Y^* = (\tilde{T} \leq t, \Delta = 1)$. For example, this could represent observed death by year 5, which would be denoted $Y^* = (\tilde{T} \leq 5 \text{ years}, \Delta = 1)$. The observed data structure for cases is then conditional on $(Y^* = 1)$. It is important to stress that the definition of a case $(Y^* = 1)$ in a nested case-control study within a right censored data structure is therefore very different than without right censoring, and accounting for this difference is not trivial. This distinction, and right censoring in general, is often overlooked in nested case-control study designs. The definition of q_0 is now $q_0 = P_0^*(\tilde{T} \leq t, \Delta = 1)$. Thus, by design we let $P_0^*(\delta = 1 \mid Y^* = 1) = 1$ and $P_0^*(\delta = 1 \mid Y^* = 0) = p$ and assume the CAR assumption $\Pi(O^*) \equiv P_0^*(\delta = 1 \mid O^*) = P_0^*(\delta = 1 \mid O_2^*)$.

Suppose we wish to compute a targeted maximum likelihood estimator for O^* of a parameter ψ_0, for example $\psi_0 = P_0^*(T_1 \leq 5 \text{ years}) - P_0^*(T_0 \leq 5 \text{ years})$, where $T_1 = (T \mid A = 1, W)$ and $T_0 = (T \mid A = 0, W)$. Thus we note that occurrence of disease conditioned upon in the case-control sampling does not need to be an outcome of interest. Targeted maximum likelihood estimators can handle both confounding as well as right censoring. To handle the right censoring, one might make use of censoring weights $\Delta / \bar{G}(\cdot)$, where $\bar{G}(\cdot)$ is the censoring mechanism, which can be estimated efficiently with a Kaplan-Meier curve (van der Laan and Rubin, 2007). Now suppose A is expensive to measure and can only be collected in a subsample of O^*. A nested case-control study might be performed. We can then implement a case-control

weighted targeted maximum likelihood estimator, as discussed in Section 3.3, with weights implied by $q_0 = P_0^*(\tilde{T} \leq 5 \text{ years}, \Delta = 1)$ in addition to the censoring weights. While simple to implement, this estimator is not a full TMLE due to the ad hoc IPCW weighting. Thus the case-control weighted IPCW TMLE is defined as the TMLE estimator for the full data structure weighting each observation $(W_i, A_i, \tilde{T}_i, \Delta_i, Y_i^*)$ with $\frac{\Delta_i q_0}{\bar{G}(\tilde{T}_i|A_i,W_i)}$ if $(Y_i^* = 1)$ and each of J corresponding control observations receive weight $\frac{\Delta_i(1-q_0)\frac{1}{J}}{\bar{G}(\tilde{T}_i|A_i,W_i)}$ if$(Y_i^* = 0)$.

An additional approach includes the use of the targeted maximum likelihood estimator presented in Moore and van der Laan (2009). This estimator involves first estimating a hazard of T given (A, W), expressing this hazard fit as a logistic regression or multiplicative intensity, and subsequently adding a time dependent covariate $h(t, A, W)$ as an epsilon extension. The epsilon coefficient in front of the clever covariate is fitted with standard logistic regression or Cox proportional hazards software, treating the initial hazard as an offset. This updating process of the conditional hazard is iterated until convergence. Once this updated hazard fit is determined with this iterative targeted maximum likelihood algorithm, one evaluates the conditional survival functions $S_{T|A=1,W}(5 \text{ years})$ and $S_{T|A=0,W}(5 \text{ years})$ and averages over W with respect to the empirical distribuiton of W. This is now the targeted maximum likelihood estimator of ψ_0, which needs to be case-control weighted by giving each observation with $(Y^* = 1)$ a weight q_0 and each control observation with $(Y^* = 0)$ a weight $(1 - q_0)\frac{1}{J}$. Note that this means each step in the above described TMLE algorithm, including the initial hazard estimation, needs to be case-control weighted.

5 Effect Modification

Nested case-control studies within clinical trials are becoming increasingly popular when researchers are interested in effect modification (Rothman and Greenland, 1998; Essebag et al., 2003, 2005; Polley and van der Laan, 2009). This is of particular importance when the patient characteristic that may modify the treatment effect is difficult or expensive to measure (Vittinghoff and Bauer, 2006). The Women's Health Initiative is an example of a well known study where the investigators' effect modification research question led to a nested case-control study design within a randomized controlled trial (Prentice and Qi, 2006). Researchers were interested in studying SNPs associated with coronary heart disease, stroke and breast cancer and hormone treatments in

their placebo controlled combined hormone trial cohort of over 16,000 women.

Suppose that within a randomized controlled trial we are interested in studying the effect modification of a particular patient characteristic, denoted W_i. The randomized controlled trial was designed with two treatment arms, $A \in \{0,1\}$, where probability of assignment was $\pi = 0.5$. The disease outcome was binary $Y \in \{0,1\}$ and the parameter:

$$\psi_0 \equiv E_0^*\{E_0^*(Y \mid A=1,W) - E_0^*(Y \mid A=0,W)\}$$

can be used to determine the average treatment effect. W indicates a multi-dimensional covariate $W = (W_i : i = 1,\ldots,m)$. However, our parameter of interest was an effect modification parameter. It represents the effect modification between $W_i \sim Bernoulli(\gamma = 0.5)$ and the treatment on the disease, while adjusting for the variables $W_{(-i)}$. This parameter of interest can be expressed:

$$\begin{aligned}\tilde{\psi}_0 &\equiv E_0^*\{[E_0^*(Y \mid A=1,A^*=1,W_{(-i)}) - E_0^*(Y \mid A=0,A^*=1,W_{(-i)})] \\ &- [E_0^*(Y \mid A=1,A^*=0,W_{(-i)}) - E_0^*(Y \mid A=0,A^*=0,W_{(-i)})]\},\end{aligned}$$

which can be written as:

$$\tilde{\psi}_0 \equiv E_0^*\{E_0^*(Z \mid A^*=1,W_{(-i)}) - E_0^*(Z \mid A^*=0,W_{(-i)})\}$$

since $\pi = 0.5$, where $Z = Y(A-(1-A))$, $A^* = W_i$, and $W_{(-i)}$ are the covariates that do not include W_i (van der Laan, 2006; Polley and van der Laan, 2009). The value of Z takes on three values, which follow a multinomial distribution:

$$Z = \begin{cases} +1 & \text{if } Y=1 \text{ and } A=1 \\ 0 & \text{if } Y=0 \\ -1 & \text{if } Y=1 \text{ and } A=0. \end{cases}$$

The effect of A^* on Z, adjusted for all other covariates $W_{(-i)}$, the parameter $\tilde{\psi}_0$, can be estimated with targeted maximum likelihood estimation. This effect estimate can be considered a causal effect modifier, if one is willing to make the assumptions discussed in Section 3.2. Now suppose A^* can only be measured in stored blood products that were collected at the beginning of the trial, and the analysis of the stored blood products in the entire trial would be prohibitively expensive. A nested case-control design would then be a natural design to study the effect modification of A^* on Z. Suppose the full data structure was defined as $O^* = (W_{(-i)}, A^*, A, Y)$. Our observed missing data structure of the nested case-control sample would then be $O = (W_{(-i)}, \delta, \delta A^*, A, Y)$. An estimate of q_0 would come from the full data, the complete randomized controlled trial.

6 Safety Analysis

Maintainers of large comprehensive databases that include adverse events, such as the General Practice Research Database (GRPD) and The Health Improvement Network (THIN), often require researchers to pay for access to the data. Cost is based on a number of factors, but almost always increases as the number of subjects requested increases. Analysis of the entire cohort of data would be cost prohibitive. Thus, nested case-control studies are also a natural design for studies of safety with pharmaceutical drugs, and our case-control methodology has the potential to provide novel insight. Recent drug safety failures (e.g., Baycol, Vioxx, Ortho Evra, and Rezulin) have led to serious side effects and deaths in users. Additional post-market evaluation tools are necessary for detecting true adverse effects among the large number of reports of side effects and adverse outcomes stored in reporting databases, which are most commonly analyzed with logistic regression, producing only conditional estimates of the odds ratio (e.g., Yang et al. (2006)). In combination with the appropriate handling of multiple testing issues, Case-Control Weighted TMLE in nested case-control studies can play an important role in the detection of true adverse events. We highlight that these are scenarios where we only have data on the case-control observations. For example, if $O^* = (W, A, Y)$, then $O = (\delta, \delta(W, A), Y)$. Thus, our estimator is maximally efficient and very appropriate for these types of study designs since no covariate information (e.g. W) on the non-case-control observations is discarded.

7 SPPARCS Data Analysis & Simulations

The National Institute of Aging funded Study of Physical Performance and Age-Related Changes in Sonomans (SPPARCS) is a population-based, census-sampled, study of the epidemiology of aging and health. Participants of this longitudinal cohort were recruited if they were aged 54 years and over and were residents of Sonoma, CA or surrounding areas. Study recruitment of 2092 persons occurred between May 1993 and December 1994 and follow-up continued for approximately 10 years. One area of particular research interest for this data has been the effect of vigorous leisure-time physical activity (LTPA) on mortality in the elderly, which has been studied in a previous collaboration (Bembom and van der Laan, 2008) using marginal structural models. The data structure $O^* = (W, A, Y)$, where $Y = I(T \leq 5 \text{ years})$, T is time to the event death, A is a binary categorization of LTPA, and W are potential confounders. These variables are further defined in Table 1. Of note is the

Table 1: **SPPARCS Variables.**

Variable		Description
Y		Death occurring within 5 years of baseline.
A		LTPA score $\geq$ 22.5 METs at baseline.‡
	HEALTH.EX	Health self-rated as "excellent."
	HEALTH.FAIR	Health self-rated as "fair."
	HEALTH.POOR	Health self-rated as "poor."
	SMOKE.CURR	Current smoker.
	SMOKE.EX	Former smoker.
W	*CARDIAC*	Cardiac event prior to baseline.
	CHRONIC	Chronic health condition at baseline.
	AGE.1	$x \leq 60$ years old.
	AGE.2	$60 < x \leq 70$ years old.
	AGE.4	$80 < x \leq 90$ years old.
	AGE.5	$x > 90$ years old.
	FEMALE	Female.

‡ LTPA is calculated from answers to a detailed questionnaire where performed vigorous physical activities are assigned standardized intensity values in metabolic equivalents (METs). The recommended level of energy expenditure for the elderly is 22.5 METs.

lack of any right censoring in this longitudinal cohort. The outcome (death within or at five years after baseline interview) and date of death was recorded for each subject. This information was available from a variety of sources, including death certificates. Our parameter of interest is the risk difference $\psi_0 = E_0^*(Y_1) - E_0^*(Y_0)$, the average treatment effect of LTPA on mortality five years after baseline interview.

The cohort was reduced to a size of $n = 2066$, as 26 subjects were missing LTPA values and/or self-rated health score (1.2% missing data). The estimated value for q_0 from the cohort was $q_{0n} = 0.130$, and the number of cases in the cohort sample was $nC = 269$. The variables used in our analysis are defined in Table 1. TMLE was performed on the full cohort sample, and the results are displayed in Table 2. Within TMLE, the machine learning Deletion/Substitution/Addition (DSA) algorithm was used to obtain estimates of $Q_0^*(A, W)$ and $g_0^*(A \mid W)$ since the functional form of the data was unknown. Our estimated parameter of interest is highly significant, and indicates that physical activity at or above recommended levels decreases five-year mortality risk in this population by 5.4%. See Table 2.

7.1 SPPARCS Simulations

We used this longitudinal cohort study to simulate nested case-control study designs where an estimate of the prevalence probability for the weights is obtained from the full cohort. For example, let us say that our full data structure $O^* = (W, A, Y)$ and observed data $O = (\delta, \delta O_1^*, O_2^*)$, where $O^* = (O_1^*, O_2^*)$, are defined by the variables in Table 1. Since this nested case-control study is simulated inside a cohort with exposure and covariate information on all controls, let us also say we set $O_1^* = (A, W)$, and $O_2^* = Y$. The SPPARCS variables W, A, and Y continue to be defined by those described in Table 1. Members of the case-control sample are denoted with $\delta = 1$. The likelihood of a single observation is then written as:

$$dP_0^*(O) \quad = \quad \{dP_0^*(W)dP_0^*(A \mid W)dP_0^*(Y \mid A, W)\}^{\delta} dP_0^*(Y)^{1-\delta}.$$

Since $O_2^* = Y$, the missing data structure ignores those individuals with $\delta = 0$, except for the purpose of estimating $P_0^*(Y = 1)$.

7.1.1 Nested Case-Control Simulations

In order to form a control sample from the SPPARCS cohort for the nested case-control design, individuals were randomly sampled from among those still alive five years from baseline interview, and assigned the value $\delta = 1$. This was a simplified approach compared to an incidence-density design where individuals are sampled from those still at risk of death at the time a case becomes a case. Sampling was performed at various sample sizes relative to the number of cases ($2nC$, $3nC$, and $4nC$). The empirical values for p in $\Pi(O^*) \equiv P_0^*(\delta = 1 \mid O^*) = P_0^*(\delta = 1 \mid O_2^*) = P_0^*(\delta = 1 \mid Y)$, with $P_0^*(\delta = 1 \mid Y = 1) = 1$ and $P_0^*(\delta = 1 \mid Y = 0) = p$, were 0.299, 0.446, and 0.608 for the three sample sizes. Non-cases that were not sampled were assigned the value $\delta = 0$. All cases were assigned $\delta = 1$.

The cohort was then resampled 1000 times. In each of the 1000 cohort resamples, one nested case-control study was extracted; those individuals with ($\delta = 1$), allowing for ties. A simulation design such as this was also used in Bureau et al. (2008). The estimated values of q_0 for use in the case-control weights for the nested case-control samples were taken from their respective cohort resample. Case-Control Weighted TMLE was performed on each of the 1000 nested case-control samples and TMLE was performed on the cohort samples. The DSA algorithm was used to obtain estimates of $Q_0^*(A, W)$ and $g_0^*(A \mid W)$ since the functional form of the data was unknown. The relative efficiency of the nested case-control parameters are compared to the cohort

parameter in Table 3, as well as average values for the parameter of interest. Relative efficiency of the nested case-control design improves as the number of controls increases. With an average of 4 controls per case (approximately 1076 of the 1797 available non-case subjects), the relative efficiency of the nested case-control design reached 78.9%.

7.1.2 Nested Case-Control Simulations with Right Censoring

For our simulations with right censored data, we generated an uninformative uniform censoring variable C, which led to 30.8% censored data in the full cohort data $O^* = (W, A, \tilde{T}, \Delta, Y^*)$. The definitions for $\tilde{T}, \Delta$, and Y^* are as described in Section 4, with W, A, and Y described in Table 1. The estimated value for q_0 from the cohort was $q_{0n} = 0.110$, and the number of cases in the cohort sample, defined by $Y^* = (\tilde{T} \leq 5 \text{ years}, \Delta = 1) = 1$, was $nC = 229$. Controls were sampled from the cohort from among those subjects who had $Y^* = 0$. The observed data for the nested case-control sample was defined as: $O = (\delta, \delta(W, A), \tilde{T}, \Delta, Y^*)$. Sampling was performed at various sample sizes relative to the number of cases as in the previous simulation, and the cohort was then resampled 1000 times. In each of the 1000 cohort resamples, one nested case-control study was extracted; those individuals with $(\delta = 1)$, allowing for ties. Values for p were 0.249, 0.371, and 0.494 for the three sample sizes. The cohort was analyzed with TMLE using IPCW weights defined as: $w_{IPCW} = \frac{I(C > \min(T,5))}{\bar{G}(\min(T,5))}$, where $\bar{G}(\cdot)$ is the censoring mechanism. The censoring mechanism can be estimated efficiently with a Kaplan-Meier curve (van der Laan and Rubin, 2007). The nested case-control samples were analyzed in a similar fashion, although we now also use IPCW weights and case-control weights in Case-Control Weighted TMLE. The relative efficiency of the nested case-control parameters are compared to the cohort in Table 4, as well as average values for the parameter of interest. Relative efficiency of the nested case-control design improves as the number of controls increases, although the nested case-control design does not reach the same high level of efficiency with 4 controls per case as our previous simulation without right censoring.

Table 2: **SPPARCS Cohort Results.** TMLE was performed on the SPPARCS cohort. Sample size was 2066, with 269 deaths five years from baseline interview and 1797 non-deaths. RD is Risk Difference, SE is Standard Error, and P is P-value.

	Estimate	SE	P
RD	-0.054	0.012	< 0.001

Table 3: **SPPARCS Simulated Nested Case-Control Results.** Case-Control Weighted TMLE was performed on the nested case-control samples, and TMLE was performed on the cohort samples. RD is Risk Difference, SE is Standard Error, RE is Relative Efficiency Compared to Cohort RD, $nC = 269$ is number of cases, and nCo is number of controls.

	Sample Size	Estimate	RE
Cohort RD	2,066	-0.055	1.000
Case-Control RD	$nCo = 2nC$	-0.101	0.319
	$nCo = 3nC$	-0.056	0.567
	$nCo = 4nC$	-0.051	0.789

Table 4: **SPPARCS Simulated Nested Case-Control Results with Right Censoring.** Case-Control Weighted IPCW TMLE was performed on the nested case-control samples, and IPCW TMLE was performed on the cohort samples. RD is Risk Difference, SE is Standard Error, RE is Relative Efficiency Compared to Cohort RD, $nC = 229$ is number of cases, and nCo is number of controls.

	Sample Size	Estimate	RE
Cohort RD	2,066	-0.064	1.000
Case-Control RD	$nCo = 2nC$	-0.040	0.270
	$nCo = 3nC$	-0.040	0.310
	$nCo = 4nC$	-0.057	0.440

8 Additional Simulation Studies

8.1 Simulated Cohort

In the SPPARCS data simulations, we did not know the true value of the parameter of interest. It is therefore important to have a completely objective way of defining the truth, and to then assess the performance of our estimator with respect to the truth. Therefore, we repeat the exact same simulation study, but now from a population we fully understand, as we know the value of the true ψ. The cohort was sampled from the target population of 1,000,000 individuals. We simulated a 5-dimensional covariate $W = (W_i : i = 1, \ldots, 5)$, a binary exposure A, and indicator Y, where 1 indicated disease (or in the case of the SPPARCS data, death by 5 years from baseline interview). These variables were generated according to the following rules:

$$W_i \sim U(0,1)$$

$$g_0^*(A \mid W) = \frac{1}{1+\exp(-(W_1+W_2+W_3+W_4))}$$

$$Q_0^*(A,W) = \frac{1}{1+\exp(-(-A-4W_1+AW_1-1.5W_2+\sin(W_5)))}.$$

The true value for the risk difference was $RD = -0.061$, and the true value for q_0 was $q_0 = 0.133$. One cohort sample was taken with 2,066 individuals, and the estimated value of q_0 taken from the cohort was $q_{0n} = 0.143$. The number of cases in the cohort sample was $nC = 296$. Controls were randomly sampled from among the non-cases in the original cohort at various sample sizes relative to the number of cases ($2nC$, $3nC$, and $4nC$), and assigned the value $\delta = 1$. Non-cases that were not sampled were assigned the value $\delta = 0$. The values for p were 0.330, 0.506, and 0.674 for the three sample sizes. All cases were assigned $\delta = 1$.

The cohort was resampled 1000 times. In each of the 1000 cohort resamples, one nested case-control study was extracted; those individuals with ($\delta = 1$), allowing for ties. Weights for the case-control samples were taken from their respective cohort resample. Case-Control Weighted TMLE was performed on each of the 1000 nested case-control samples and TMLE was performed on the cohort samples. Logistic regression was used to estimate $Q_0^*(A,W)$ and $g_0^*(A \mid W)$ since the functional form was known. The relative efficiency of the nested case-control parameters are compared to the cohort in Table 5, as well as average values for the parameter of interest. As before, relative efficiency of the nested case-control design improves as the number of controls increases.

With an average of 4 controls per case, the nested design reaches a relative efficiency of 78.4%. Bias results can be seen in Figure 1.

8.2 Simulated Clinical Trial

As previously discussed, nested case-control studies within clinical trials are becoming increasingly common when researchers are interested in effect modification. Thus, we provide an additional illustrative example of our methods for this research question. The simulated target population contained 1,000,000 individuals with covariates W. For the clinical trial, 10,000 were sampled and assigned a treatment A. The outcome of disease was assigned with $Y = 1/(1+\exp(-(3A-4W_1+W_3-12W_4-2W_5+2A\sin(W_3))))$. Of the 10,000 subjects, 647 individuals developed disease (6.47%). The value of the effect modification parameter of interest in the full trial was $\tilde{\psi}_0 = E_0^*\{E_0^*(Z \mid A^* = 1, W_{(-i)}) - E_0^*(Z \mid A^* = 0, W_{(-i)})\} = 0.016$. The full data in the randomized controlled trial cohort was analyzed with TMLE.

However, suppose that the effect modifier of interest, $W_3 \equiv A^*$, could only be measured in stored blood products, which is a very expensive process. Therefore, we could not measure $\tilde{\psi}_0$, as discussed in Section 5, in the entire trial and chose a nested case-control design. In order to simulate a nested case-control study within our simulated clinical trial data, controls were randomly sampled from among the non-cases in the original cohort at various sample sizes relative to the number of cases ($2nC$, $3nC$, $4nC$, and $5nC$), and assigned $\delta = 1$. Non-cases that were not sampled were assigned $\delta = 0$. The values for p were 0.141, 0.210, 0.280, and 0.350 for the four sample sizes. All subjects with $Y = 1$ were assigned $\delta = 1$. The resampling procedure was the same as our previous simulated designs. Case-Control Weighted TMLE was used to analyze the nested case-control samples. Multinomial regression was used with main terms to estimate $Q_0^*(A^*, W)$, and this represents a misspecified model. Due to the double robustness of the TMLE and Case-Control Weighted TMLE procedures, the estimates of the parameter of interest are consistent even when $Q_0^*(A^*, W)$ or $g_0^*(A^* \mid W)$ is misspecified. The values for $g_0^*(A^* \mid W)$ were known since it was a randomized controlled trial. Results are displayed in Table 6. The relative efficiency of the nested case-control design improves as the number of controls increases, and with 38.8% of the total trial participants we reach an efficiency of 86.4%.

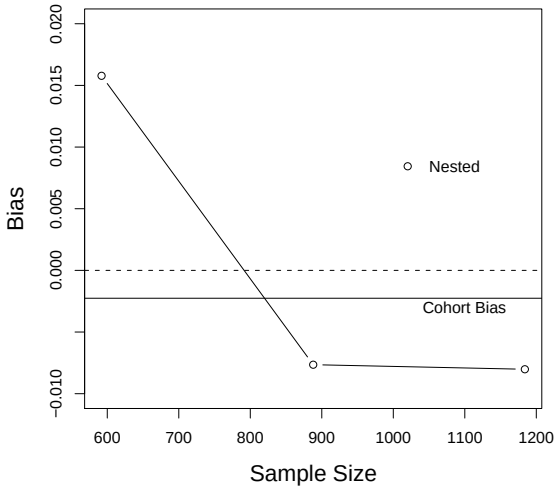


Figure 1: **Simulation Data Nested Case-Control – Bias Results for the Risk Difference.**

Table 5: **Simulation Data Nested Case-Control Results.** Case-Control Weighted TMLE was performed on the nested case-control samples and TMLE was performed on the cohort samples. RD is Risk Difference, SE is Standard Error, RE is Relative Efficiency Compared to Cohort RD, $nC = 296$ is number of cases, and nCo is number of controls.

	Sample Size	Estimate	RE
Cohort RD	2,066	-0.063	1.000
	$nCo = 2nC$	-0.045	0.411
Case-Control RD	$nCo = 3nC$	-0.068	0.725
	$nCo = 4nC$	-0.069	0.788

Table 6: **Randomized Controlled Trial Simulation Data Nested Case-Control Results.** Case-Control Weighted TMLE was performed on the nested case-control samples and TMLE was performed on the full trial samples. SE is Standard Error, RE is Relative Efficiency Compared to Cohort RD, $nC = 647$ is number of cases, and nCo is number of controls.

	Sample Size	Estimate	RE
Full Trial $\tilde{\psi}$	10,000	0.016	1.000
Case-Control $\tilde{\psi}$	$nCo = 2nC$	0.024	0.142
	$nCo = 3nC$	0.022	0.253
	$nCo = 4nC$	0.019	0.517
	$nCo = 5nC$	0.016	0.864

9 Discussion

Nested designs have the potential to significantly reduce the costs associated with collecting data on the full cohort with only minimal losses in efficiency (Ernster, 1994; Rothman and Greenland, 1998; Hak et al., 2004; Vittinghoff and Bauer, 2006). Our simulated nested case-control studies within the SPPARCS data demonstrated 78.9% efficiency with an average of 4 controls per case. We had 78.4% efficiency in our simulated nested case-control studies within a simulated cohort, again with an average of 4 controls per case. These results coincided with the conclusions of Ury (1975), which noted that as a general rule, 4 controls per case yields a relative efficiency of 80.0%. Our nested case-control simulations with right censoring within the SPPARCS data also demonstrated that methods for right censoring can be incorporated into the Case-Control Weighted TMLE procedure. In general, our case-control methodology can be used in conjunction with procedures that handle censoring, missingness, measurement error, and other persistent issues found in public health and medicine. We also demonstrated the use of Case-Control Weighted TMLE for nested case-control study designs within randomized controlled trials when interested in an effect modification research question. With less than 40% of the trial subjects, we reached an efficiency of 86.4% compared to the full trial.

The extension of our Case-Control Weighted TMLE methodology to nested case-control study designs provides a double robust locally efficient estimation procedure for marginal causal effects and variable importance measures in nested designs. We showed that both the case-control weighted targeted maximum likelihood estimator and the IPCW estimator are targeted maximum

likelihood estimators for the missing data structure $(\delta, \delta(W, A), Y)$, and are thus locally efficient procedures for that data. For appropriate inference (e.g. construction of standard errors), however, the IPCW efficient influence curve must be implemented, or an appropriate resampling procedure such as bootstrapping. With the increase in popularity of nested case-control study designs in longitudinal cohorts and randomized controlled trials, the extension of our Case-Control Weighted TMLE procedure provides an additional tool to yield unique biological and public health discovery.

References

J.A. Anderson. Separate sample logistic discrimination. *Biometrika*, 59:19–35, 1972.

W.E. Barlow, L. Ichikawa, D. Rosner, and S. Izumi. Analysis of case-cohort designs. *J Clin Epidemiol*, 52(12):1165–1172, 1999.

O. Bembom and M.J. van der Laan. Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators. *Technical Report 230, Division of Biostatistics, University of California, Berkeley*, 2008.

O. Borgan and B. Langholz. Nonparametric estimation of relative mortality from nested case-control studies. *Biometrics*, pages 593–602, 1993.

N.E. Breslow and K.C. Cain. Logistic regression for two-stage case-control data. *Biometrika*, 75(1):11–20, 1988.

N.E. Breslow, J.H. Lubin, and P. Marek. Multiplicative models and cohort analysis. *J Am Stat Assoc*, 78:1–12, 1983.

A. Bureau, M.S. Diallo, J.M. Ordovas, and L.A. Cupples. Estimating interaction between genetic and environmental risk factors: Efficiency of sampling designs within a cohort. *Epidemiology*, 19(1):83–93, 2008.

V.L. Ernster. Nested case-control studies. *Prev Med*, 23(5):587–590, 1994.

V. Essebag, J. Genest Jr., S. Suissa, and L. Pilote. The nested case-control study in cardiology. *American Heart Journal*, 146(4):581–590, 2003.

V. Essebag, R.W. Platt, M. Abrahamowicz, and L. Pilote. Comparison of nested case-control and survival analysis methodologies for analysis of time-dependent exposure. *BMC Medical Research Methodology*, 5(5), 2005.

W.D. Flanders and S. Greenland. Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine*, 10(5), 1991.

S. Greenland. Multivariate estimation of exposure-specific incidence from case-control studies. *J Chron Dis*, 34:445–453, 1981.

E. Hak, F. Wei, D.E. Grobbee, and K.L. Nichol. A nested case-control study of influenza vaccination was a cost-effective alternative to a full cohort analysis. *J Clin Epidemiol*, 57(9):875–880, 2004.

L.L. Kupper, A.J. McMichael, and R. Spirtas. A hybrid epidemiologic study design useful in estimating relative risk. *J Am Stat Assoc*, (70):524–528, 1975.

F.D.K. Liddell, J.C. McDonald, and D.C. Thomas. Methods of cohort analysis: appraisal by application to asbestos mining. *J R Stat Soc Ser A*, (140):469–491, 1977.

J.H. Lubin. Extensions of analytic methods for nested and population-based incident case-control studies. *J Chronic Dis*, 39(5):379–388, 1986.

N. Mantel. Synthetic retrospective studies and related topics. *Biometrics*, 29 (3):479–486, 1973.

K. Moore and M.J. van der Laan. Application of time-to-event methods in the assessment of safety in clinical trials. In K.E. Peace, editor, *Design and Analysis of Clinical Trials with Time-to-Event Endpoints*. Chapman & Hall/CRC Biostatistics Series, 2009.

A.P. Morise, G.A. Diamon, R. Detrano, M. Bobbio, and Erdogan Gunel. The effect of disease-prevalence adjustments on the accuracy of a logistic prediction model. *Med Decis Making*, 16:133–142, 1996.

E.C. Polley and M.J. van der Laan. Selecting optimal treatments based on predictive factors. In K.E. Peace, editor, *Design and Analysis of Clinical Trials with Time-to-Event Endpoints.* Chapman & Hall/CRC Biostatistics Series, 2009.

R.L. Prentice and N.E. Breslow. Retrospective studies and failure time models. *Biometrika*, 65(1):153–158, 1978.

R.L. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66:403–411, 1979.

R.L. Prentice and L. Qi. Aspects of the design and analysis of high-dimensional snp studies for disease risk estimation. *Biostatistics*, 7(3):339–354, 2006.

J.M. Robins. [choice as an alternative to control in observational studies]: Comment. *Statistical Science*, 14(3):281–293, 1999.

J.M. Robins, A. Rotnitzky, and L.P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

S. Rose and M.J. van der Laan. Simple optimal weighting of cases and controls in case-control studies. *The International Journal of Biostatistics*, 4(1): Article 19, 2008.

S. Rose and M.J. van der Laan. Why match? Investigating matched case-control study designs with causal effect estimation. *The International Journal of Biostatistics*, 5(1):Article 1, 2009.

K. Rothman and S. Greenland. *Modern Epidemiology*. Lippincott, Williams and Wilkins, Philadelphia, PA, 2nd edition, 1998.

S.O. Samuelsen. A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika*, 84(2):379–394, 1997.

M. Szklo and F.J. Nieto. *Epidemiology: Beyond the Basics*. Jones & Bartlett Publishers, Boston, MA, 2nd edition, 1999.

D.C. Thomas. Addendum to: "Methods of cohort analysis: appraisal by application to asbestos mining" by F.D.K. Liddell and J.C. McDonald and D.C. Thomas. *J R Stat Soc Ser A*, (140):469–491, 1977.

H.K. Ury. Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data. *Biometrics*, 31(3):643–649, 1975.

M.J. van der Laan. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1):Article 2, 2006.

M.J. van der Laan. Estimation based on case-control designs with known prevalence probability. *The International Journal of Biostatistics*, 4(1): Article 17, 2008.

M.J. van der Laan and J.M. Robins. *Unified methods for censored longitudinal data and causality*. Springer, New York, 2003.

M.J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):Article 11, 2006.

M.J. van der Laan and D. Rubin. A note on targeted maximum likelihood and right censored data. *Technical Report 226, Division of Biostatistics, University of California, Berkeley*, 2007.

E. Vittinghoff and D.C. Bauer. Case-only analysis of treatment-covariate interactions in clinical trials. *Biometrics*, 62(3):769–776, 2006.

S. Wacholder. The case-control study as data missing by design: Estimating risk differences. *Epidemiology*, 7(2):144–150, 1996.

Y.X. Yang, J.D. Lewis, S. Epstein, and D.C. Metz. Long-term proton pump inhibitor therapy and risk of hip fracture. *JAMA*, 296(24):2947–2953, 2006.

Chapter 9

Time-to-Event Outcomes and Censored Data

9.1 *A Note on Targeted Maximum Likelihood and Right Censored Data*

The following article appears as it was originally published on the University of California, Berkeley Division of Biostatistics Working Paper Series website in 2007, http://www.bepress.com/ucbbiostat/paper226/.

A Note on Targeted Maximum Likelihood and Right Censored Data

Mark J. van der Laan and Daniel B. Rubin

A popular way to estimate an unknown parameter is with substitution, or evaluating the parameter at a likelihood-based fit of the data generating density. In many cases, such estimators have substantial bias and can fail to converge at the parametric rate. van der Laan and Rubin (2006) introduced targeted maximum likelihood learning, removing these shackles from substitution estimators, which were made in full agreement with the locally efficient estimating equation procedures as presented in Robins and Rotnitzky (1992) and van der Laan and Robins (2003). This note illustrates how targeted maximum likelihood can be applied in right censored data structures. In particular, we show that when an initial substitution estimator is based on a Cox proportional hazards model, the targeted likelihood algorithm can be implemented by iteratively adding an appropriate time-dependent covariate.

1 Introduction

Suppose we observe a sample $\{O_i\}_{i=1}^n$ of independent and identically distributed observations, for

$$O = (W, \Delta = I(T \leq C), \tilde{T} = \min(T, C)) \sim P \in \mathcal{M}. \quad (1)$$

Here W is a vector of baseline covariates, T is a survival time, C is a censoring time, Δ is an indicator of censoring, P is the data generating distribution, and the statistical model $\mathcal{M}$ is a family of data generating distributions containing P. We will make the usual assumption that

$$\{T \perp C|W\}, \quad (2)$$

meaning survival and censoring times are conditionally independent given the baseline covariates. The log likelihood for a single observation can be written as

$$\begin{aligned} dP(w, \delta, \tilde{t}) &= dP(W = w) \\ &\times [dP(T = \tilde{t}|W = w)P(C \geq \tilde{t}|W = w)]^{\delta} \\ &\times [P(T > \tilde{t}|W = w)dP(C = \tilde{t}|W = w)]^{1-\delta}. \end{aligned} \quad (3)$$

The full data, which would have liked to observe, but could not be completely measured because of censoring, consists of the baseline covariates and survival times $\{X_i\}_{i=1}^n = \{W_i, T_i\} \sim F$. We can write $P = P_{F,G}$, for $G(\cdot|W)$ denoting the conditional cumulative distribution function of the censoring time. This note applies to general scenarios where the goal is to estimate a smooth (pathwise differentiable) Euclidean parameter $\mu(F) \in \mathbb{R}^k$, representing some feature of the full data distribution.

An example of such a parameter is simply the marginal survival probability at a fixed time t,

$$\mu(P_{F,G}) = \mu(F) = \bar{F}(t) = P(T > t). \quad (4)$$

Note that if the stronger unconditional independence assumption $\{T \perp C\}$ doesn't hold, the (1958) Kaplan-Meier estimator might not necessarily be consistent. Even if Kaplan-Meier assumptions aren't violated, the presence of informative baseline covariates make efficiency gains possible. When a randomized treatment $A \in \{0, 1\}$ is assigned at baseline, another important parameter could be the risk difference

$$\mu(F) = \bar{F}(t|A = 1) - \bar{F}(t|A = 0) = P(T > t|A = 1) - P(T > t|A = 0). \quad (5)$$

Additionally, interest might lie in regression parameters such as

$$\mu(F) = (\beta_0(F), \beta(F)) = \text{argmin}_{\beta_0,\beta} E_F|\log(T) - \beta_0 - \beta^T W|^2. \quad (6)$$

Note that $\mu(F)$ can here be defined without assuming an accelerated failure time model actually holds. It is simply a coefficient vector giving the best linear predictor of log survival from baseline covariates.

For many parameters of interest, the prevailing estimation technique is to first fit the data generating distribution $P \in \mathcal{M}$ with some $\hat{P} \in \mathcal{M}$ according to maximizing likelihood over a submodel $\mathcal{M}_0 \subset \mathcal{M}$, and then forming the substitution estimator $\hat{\mu} = \mu(\hat{P})$. This note will focus on how to proceed when initially considering a substitution estimator based on the ubiquitous proportional hazards model introduced in Cox (1973). Unfortunately, substitution estimators often have poor performance. As discussed in Robins and Ritov (1997), they can be heavily biased because the choice of $\hat{P} \in \mathcal{M}$ was made without regard to the parameter of interest. Such estimators can be inconsistent, or lead to arbitrarily bad rates of convergence, while simpler schemes can sometimes guarantee the parametric $n^{-1/2}$ rate.

For example, Robins and Rotnitzky (2005) review inverse probability of censoring weighted (IPCW) estimators in survival analysis, which can lead to $\sqrt{n}$-consistent, asymptotically linear estimators if the censoring mechanism $\bar{G}(\cdot|W)$ can be well approximated. Frequently censoring is caused by study termination, and the censoring time is independent of the survival time and baseline covariates, in which case $\bar{G}(\cdot|W) = \bar{G}(\cdot)$ can be efficiently estimated with the Kaplan-Meier curve.

Suppose $D_{\text{Full}}(W, T|F) = D_{\text{Full}}(W, T|\mu(F), \eta(F)) : (W, T) \to \mathbb{R}^k$ is an estimating function for μ we could use with access to the full data $\{X_i = (W_i, T_i)\}_{i=1}^n$. That is, suppose $E_F[D_{\text{Full}}(W, T|\mu, \eta(F))] = 0$ at $\mu = \mu(F)$, and that with no censoring we could reliably estimate $\mu(F)$ with the solution to

$$0 = \frac{1}{n} \sum_{i=1}^{n} D_{\text{Full}}(W_i, T_i|\mu, \eta_n). \quad (7)$$

Here the left side is the zero vector in $\mathbb{R}^k$, and η_n is an estimator of the nuisance parameter $\eta(F)$. For the three parameters given by (4), (5), and (6), respective full data estimating equations could be

$$D_{\text{Full}}(W, T|F) = I(T > t) - \mu(F) \quad (8)$$

$$D_{\text{Full}}(W, T|F) = I(T > t)\left(\frac{A}{P(A=1)} - \frac{1-A}{P(A=0)}\right) - \mu(F) \quad (9)$$

$$D_{\text{Full}}(W, T|F) = W(\log(T) - \beta_0 - \beta^T W) \text{ recalling } \mu(F) = [\beta_0, \beta]^T. \quad (10)$$

Inverse probability of censoring weighted estimation maps the full data estimating function into

$$D_{\text{IPCW}}(O|P) = D_{\text{IPCW}}(O|\mu, \eta(F), G) = \frac{D_{\text{Full}}(O|\mu, \eta(F))\Delta}{\bar{G}(\tilde{T}_-|W)}, \qquad (11)$$

which is a function of the observed data $O = (W, \Delta, \tilde{T})$. It is easy to verify $E_P[D_{\text{IPCW}}(O|P)] = E_F[D_{\text{Full}}(W, T|F)] = 0$. Hence, we can use it as an estimating equation for $\mu(F)$, after fitting nuisance parameters $\eta(F)$ and $G(\cdot|W)$. While simple, IPCW estimating equations are suboptimal in terms of both efficiency and robustness. We refer to van der Laan and Robins (2003) for a survey of estimating function methodology in survival analysis.

Despite advantages of the estimating function methodology outlined in this survey, likelihood based substitution estimators remain more prevalent in many applications. This could be for a variety of reasons, among them outlier concerns due to inverse weighting, computational considerations, unfamiliarity, and inertia. To remedy the situation, van der Laan and Rubin (2006) introduced targeted maximum likelihood. Given an initial fit $\hat{P}$ of the data generating distribution, the procedure iteratively updates the fit by maximizing likelihood along submodels chosen to best target the parameter of interest $\mu(F)$. The algorithm maps an initial $\hat{P}$ into a $\hat{P}^\star \in \mathcal{M}$, at which the substitution estimator $\mu(\hat{P}^\star)$ is also the solution to a well-chosen estimating equation. Hence, the resulting estimator is a familiar type of likelihood based substitution estimator, inheriting the benefits of $\sqrt{n}$-convergence, asymptotic linearity, and local efficiency implied by estimating function theory. Targeted maximum likelihood works as follows:

1. Form an initial fit $\hat{P} \in \mathcal{M}$ of the data generating distribution.

2. Create a smooth (regular) parametric submodel of $\mathcal{M}$, parametrized by an ϵ, passing through $\hat{P}$ at $\epsilon = 0$. Ensure the linear span of the score vector at $\hat{P}$ includes the efficient influence curve for parameter $\mu(P)$ at $\hat{P}$. The efficient influence curve will be discussed in the sequel, and is formally discussed in Bickel et al. (1998) and Chapter 1.4 of van der Laan and Robins (2003).

3. Estimate ϵ with maximum likelihood.

4. Define a new density estimator as the corresponding update to the original estimator $\hat{P}$.

5. Iterate steps 2-4 until convergence. Of course, the procedure can be applied without iteration, and van der Laan and Rubin (2006) argued that most bias reduction should occur in the first step.

The efficient influence curve $D(O|P) = D(O|\mu, \eta, G, F)$, or a scaled version thereof, is in a strong sense the optimal estimating equation for the parameter of interest. If the nuisance parameters on which it depends are estimated accurately, and regularity conditions are met, the estimating equation gives rise to the regular asymptotically linear estimator with the smallest possible asymptotic variance. Further, as discussed in van der Laan and Robins (2003), it has desirable robustness properties. If either the initial full data fit $\hat{F}$ is a good approximation to F, or the censoring mechanism estimate $\hat{G}$ is a good approximation to G, using the estimating function $D(O|\mu, \eta, \hat{G}, \hat{F})$ can ensure asymptotic linearity.

Suppose no special parametric or semiparametric assumptions are made on the full data model, and the $D_{\text{Full}}(W, T|F)$ given earlier would be a valid estimating equation with uncensored data. Robins and Rotnitzsky (1992) show the efficient influence curve at $P_{F,G}$ is given by a scaled version of,

$$\begin{aligned} D(O|P) &= D(O|\mu(P), \eta(P), F(P), G(P)) \qquad (12) \\ &= D_{\text{IPCW}}(O|\mu, \eta, G) + \int_t \frac{E_F[D_{\text{Full}}(W, T|\mu, \eta)|W, T > t]}{\bar{G}(\tilde{T}|W)} dM_G(t). \end{aligned}$$

Here the last term is an integral with respect to the martingale,

$$M_G(t) = I(\tilde{T} \leq t, \Delta = 0) - \int_{-\infty}^{t} I(\tilde{T} \geq s) \frac{dG(s|W)}{\bar{G}(s_-|W)}. \qquad (13)$$

Examine the targeted likelihood algorithm. Upon convergence to $\hat{P}^\star$, the relevant submodel's likelihood is maximized at $\epsilon = 0$. Hence, the score at $\epsilon = 0$ will have empirical mean zero. But from the choice of submodel, this means the efficient influence curve $D(O|\hat{P}^\star) = D(O|\mu(\hat{P}^\star), \eta(\hat{P}^\star), F(\hat{P}^\star), G(\hat{P}^\star))$ will have empirical mean zero. In other words, the substitution estimator $\mu(\hat{P}^\star)$ will solve the efficient influence curve estimating equation, based on plug-in estimators for the curve's nuisance parameters.

This note is devoted to showing how targeted likelihood can be implemented when the initial fit is based on Cox's proportional hazards model. The initial fit to the data generating distribution can be decomposed into fits of the baseline covariate distribution P_W, the censoring mechanism $G(|W)$ representing the conditional distribution $\mathcal{L}(C|W)$, and the conditional survival distribution $\mathcal{L}(T|W)$. As we'll mention in Section 3, it will be convenient to

use the empirical distribution placing mass $\frac{1}{n}$ on $W_1, ..., W_n$ to fit the baseline covariate distribution. As previously observed, the censoring mechanism can be fit with the Kaplan-Meier product-limit estimator if we believe censoring is independent of survival, but arbitrary initial fits can be used. Neither the baseline covariate distribution fit not the censoring mechanism fit will be updated at any step of the targeted likelihood algorithm. The Cox model is meant for estimating the conditional survival distribution. Note that the methodology can applied without necessarily believing the model holds, and targeted likelihood can allow us to consistently estimate parameters such as (4), (5), and (6) using a misspecified model. We'll consider a variant of the model assuming,

$$\Lambda(t|W) = \int_{-\infty}^{t} \frac{dF(s|W)}{\bar{F}(s_-|W)} = \Lambda_0(t)\exp(\beta^T L(W,t)). \tag{14}$$

Here $L(\cdot, W)$ is a specified function allowing multiplicative effect on conditional hazard to change with time. Coefficient vector β can be estimated by $\hat{\beta}$ through maximizing Cox's (1973) partial likelihood, while the Breslow (1974) estimator $\hat{\Lambda}_0(\cdot)$ is commonly used to fit the baseline cumulative hazard function $\Lambda_0(\cdot)$. Together, these fits determine a fit $\hat{\Lambda}(\cdot|W)$ of the conditional cumulative hazard $\Lambda(\cdot|W)$, and consequently the conditional survival distribution $\mathcal{L}(T|W)$. Taken together, $\hat{P}_W$, $\hat{G}(\cdot|W)$ and $\hat{\Lambda}(\cdot|W)$ determine the initial fit $\hat{P}$ of the data generating distribution. This is step 1 of the targeted likelihood algorithm, and it remains to be seen how $\hat{P}$ can be mapped into the $\hat{P}^\star$ providing an accurate substitution estimator for the parameter of interest.

2 Statement of Main Result

The targeted likelihood algorithm can be implemented by iteratively adding an appropriate time-dependent covariate to the Cox proportional hazards model. Letting $\hat{P}$ denote the initial data generating fit just mentioned, and $\bar{G}_n(\cdot|W) = 1 - \hat{G}(\cdot|W)$ the corresponding censoring mechanism fit, define the function

$$h(w,t|\hat{P}) = \frac{D_{\text{Full}}(w,t|\hat{P}) - E_{\hat{P}}[D_{\text{Full}}(w,T|\hat{P})|W=w, T>t]}{\bar{G}_n(t_-|w)}. \tag{15}$$

For fixed baseline cumulative hazard fit $\hat{\Lambda}(\cdot)$ and coefficient vector fit $\hat{\beta}$, consider the submodel

$$\Lambda_\epsilon(t|W) = \hat{\Lambda}_0(t)\exp(\hat{\beta}^T L(W,t) + \epsilon^T h(W,t|\hat{P})), \tag{16}$$

parametrized by $\epsilon \in \mathbb{R}^k$. Here ϵ has the same dimension as the parameter $\mu(F)$ and efficient influence curve $D(O|P)$.

Choosing $\hat{\epsilon}$ to maximize the likelihood of observed data $\{O_i\}_{i=1}^n$ corresponds to carrying out an iteration of the targeted maximum likelihood algorithm. The remainder of this note sketches the argument, without attempting to be overly formal.

When the data generating distribution fit $\hat{P}$ is updated based on the fit to this model, the procedure can be iterated. Hence, iteration corresponds to repeatedly adding a time-dependent covariate vector to an existing proportional hazards model, and using maximum likelihood to fit the associated coefficient vector while keeping everything else in the model fixed. Standard software can be used to fit ϵ via maximum likelihood, but this will require being able to evaluate covariate $h(W, t|\hat{P})$, which could be cumbersome due to the conditional expectation in its second term.

3 Sketch of Argument that Adding Covariate Implements Targeted Likelihood Algorithm

Following Bickel et. al. (1998), we can define the tangent space $T(P)$ as the closure in $L_0^2(P)$ of the linear span of all scores of regular parametric submodels of $\mathcal{M}$ through P. It is well known that if the model is nonparametric, the tangent space is saturated, meaning that $T(P) = L_0^2(P)$. It is also easy to see the tangent space can be decomposed into the three tangent spaces corresponding to scores through P fluctuating the baseline covariate distribution $\mathcal{L}(W)$, conditional survival distribution $\mathcal{L}(T|W)$, and censoring mechanism $\mathcal{L}(C|W)$. These three tangent spaces

$$\begin{aligned} T_{\mathrm{W}}(P) &= \{r(W) \in L_0^2(P) : \ E[r(W)] = 0\} && (17) \\ T_{\mathrm{F}}(P) &= \{v(O) \in L_0^2(P) : \ E[v(O)|C, W] = 0\} && (18) \\ T_{\mathrm{CAR}}(P) &= \{v(O) \in L_0^2(P) : \ E[v(O)|T, W] = 0\} && (19) \end{aligned}$$

are orthogonal, giving us the direct sum

$$T(P) = L_0^2(P) = T_{\mathrm{W}}(P) \oplus T_{\mathrm{F}}(P) \oplus T_{\mathrm{CAR}}(P) \tag{20}$$

and the decomposition of any $v(O) \in L_0^2(P)$ into

$$\begin{aligned} v(O) &= \Pi(v(\cdot)|T(P))(O) && (21) \\ &= \Pi(v(\cdot)|T_{\mathrm{W}}(P))(O) + \Pi(v(\cdot)|T_{\mathrm{F}}(P))(O) + \Pi(v(\cdot)|T_{\mathrm{CAR}}(P))(O). \end{aligned}$$

This decomposition can be applied to the efficient influence curve $D(O|P)$. To find a submodel through P with score equal to this influence curve, it is thus only necessary to find submodels varying $\mathcal{L}(W)$, $\mathcal{L}(T|W)$, $\mathcal{L}(C|W)$ that give $\Pi(D(\cdot|P)|T_{\mathrm{W}}(P))(O)$, $\Pi(D(\cdot|P)|T_{\mathrm{F}}(P))(O)$, and $\Pi(D(\cdot|P)|T_{\mathrm{CAR}}(P))(O)$ as their respective scores.

3.1 Baseline Covariate Distribution

Letting $\hat{P}_W$ denote the empirical distribution on the baseline covariates $\{W_i\}_{i=1}^n$ given as the initial fit in the previous section, and $\hat{P}$ the initial fit for the entire data generating distribution P, we can trivially define the submodel

$$dP_W^{(\delta)} = \frac{\exp(\delta\Pi(D(O|\hat{P})|T_W(\hat{P})))}{\int \exp(\delta\Pi(D(O|\hat{P})|T_W(\hat{P})))d\hat{P}_W^{(\delta)}} d\hat{P}_W. \tag{22}$$

The projection operator is given by $\Pi(v(O)|T_W(\hat{P})) = E_{\hat{P}}[v(O)|W]$, but this will not be relevant for our purposes. It can be verified that this submodel gives the desired score of $\Pi(D(O|\hat{P})|T_{\mathrm{W}}(\hat{P}))$.

In fact, the exponential family technique can always be used to define a submodel of a nonparametric model having a desired score. We could have simply used the exponential family $dP^{(\delta)}(O) \propto \exp(\delta D(O|\hat{P})d\hat{P}(O)$ for the entire data generating distribution, but targeted likelihood becomes more difficult to implement than in our Cox model formulation.

The specific choice of submodel through P_W is not at all important for the targeted likelihood procedure, so long as it gives rise to the correct score. This is because $\hat{P}_W$ is never updated from its initial empirical distribution fit, as this is the nonparametric maximum likelihood estimate (NPMLE) for P_W. Consequently, in each iteration of the targeted likelihood algorithm, the $\hat{P}_k$ to be used as a substitution estimator corresponds to using the empirical distribution baseline covariate fit.

We mean to focus attention on when the survival distribution, meaning the marginal $\mathcal{L}(T)$ or conditional $\mathcal{L}(T|W)$ law, is of primary interest, rather than the baseline covariate distribution $\mathcal{L}(W)$. If there is concern substitution estimation of $\mu(F)$ based on the empirical $\hat{P}_W$ might lead us astray, the problem would have to be reconsidered.

3.2 Censoring Mechanism

As discussed in Chapter 1.4.4 of van der Laan and Robins (2003), the efficient influence curve $D(O|P)$ is orthogonal to the tangent space T_{CAR} generated

from scores of submodels varying the censoring mechanism $\mathcal{L}(C|W)$. Hence, $\Pi(D(\cdot|P)|T_{\text{CAR}}(P)) = 0$, and we do not need to perturb the censoring mechanism from its initial fit in the targeted maximum likelihood algorithm.

3.3 Conditional Survival Time Distribution

Note from Chapter 1.4 of van der Laan and Robins (2003) that the efficient influence curve at P can be written as

$$D(O|P) = D_{\text{IPCW}}(O|P) - \Pi(D(\cdot|P)|T_{\text{CAR}}(P))(O), \tag{23}$$

and that $T_{\text{F}}(P)$ is orthogonal to $T_{\text{CAR}}(P)$. Together these facts clearly imply

$$\Pi(D(\cdot|P)|T_{\text{F}}(P)) = \Pi(D_{\text{IPCW}}(\cdot|P)|T_{\text{F}}(P)). \tag{24}$$

Thus, we only need to show the submodel through the $\mathcal{L}(T|W)$ fit in the previous section gives rise to a score equal to the IPCW estimating function's projection on tangent space $T_{\text{F}}(P)$.

Define the counting process $N(t) = I(\tilde{T} \leq t, \Delta = 1)$ jumping at an observed failure time. Recalling $\Lambda(\cdot|W)$ represents the conditional cumulative hazard function for $\mathcal{L}(T|W)$, the associated Doob-Meyer martingale is

$$M(t) = N(t) - \int_{-\infty}^{t} I(\tilde{T} \geq s) d\Lambda(s|W). \tag{25}$$

From Theorem 1.1 of van der Laan and Robins (2003), interchanging the completely symmetric $T_{\text{CAR}}(P)$ and $T_{\text{F}}(P)$, the projection operator is given by

$$\Pi(v|T_{\text{F}}(P)) = \int (E_P[v(O)|W, T = t, C \geq t] - E_P[v(O)|W, T > t, C \geq t]) dM(t). \tag{26}$$

We can apply this result with $v(O) = D_{\text{IPCW}}(O) = \frac{D_{\text{Full}}(W,T|P)\Delta}{\bar{G}(\tilde{T}_-|W)}$. Given that $\{T = t, C \geq t\}$ implies $\Delta = 1$, it is clear $E_P[v(O)|W, T = t, C \geq t] = \frac{D_{\text{Full}}(W,t|P)}{\bar{G}(t_-|W)}$. Further, it is an elementary calculation to show $E_P[v(O)|W, T > t, C \geq t]$ is equal to $E_P[D_{\text{Full}}(W, T|P)|W, T > t]/\bar{G}(t_-|W)$. Hence, the efficient influence curve $D(O|P)$ has projection on tangent space $T_{\text{F}}(P)$ of

$$\Pi(D(\cdot|P)|T_{\text{F}}(P)) = \int h(W, t|P) dM(t), \tag{27}$$

for the $h(W, t|P)$ defined in (15). However, as reviewed in Lemma 3.2 of van der Laan and Robins (2003), $\int g(W, t) dM(t)$ is simply the score at $\epsilon = 0$

of a submodel through P varying conditional cumulative hazard of $\mathcal{L}(T|W)$ through

$$\Lambda_\epsilon(t|W) = \Lambda(t|W)\exp(\epsilon^T g(W,t|P)). \tag{28}$$

Thus, the projection $\Pi(D(\cdot|\hat{P})|T_{\mathrm{F}}(\hat{P}))$ in $L_0^2(\hat{P})$ is exactly the score at $\epsilon = 0$ of the submodel (16). Recall that this was the desired result, from our decomposition of $D(O|P)$ into projections on $T_{\mathrm{W}}(P)$, $T_{\mathrm{CAR}}(P)$ and $T_{\mathrm{F}}(P)$. By adding $h(W,t|\hat{P})$ as a time-dependent covariate to a Cox model, fixing the censoring mechanism fit, and placing a submodel through the baseline covariate empirical distribution fit, we can obtain the efficient influence curve as a score. Because the baseline covariate fit will never be perturbed, targeted likelihood proceeds by iteratively updating the initial Cox model fit.

4 Discussion

In this note, we've shown how a Cox-based substitution estimator can be made to solve a locally efficient estimating equation, if appropriate covariates are added to an initial fit. Estimating equation approaches are often avoided in favor of more familiar substitution estimators, despite their theoretical advantages outlined in van der Laan and Robins (2003). By representing estimating function procedures as fits to commonplace Cox models, we hope to make the methodology more amenable. This parallels results given in van der Laan and Rubin (2006) and Moore and van der Laan (2007) demonstrating the targeted likelihood algorithm can be implemented in causal inference problems by adding covariates to linear and logistic regression models, although in those cases the algorithm was shown to converge in a single iteration.

Several serious caveats are in order. Primarily, while we've suggested how to perform targeted maximum likelihood, our exposition was hardly a formal proof. Further, van der Laan and Rubin (2006) listed several criteria to ensure convergence of the iterative algorithm, which have not been checked in this work, although we expect them to hold. Finally, while it sounds straightforward to iteratively add a time-dependent covariate to a Cox model, we have glossed over the specific details of how to implement our procedure.

Bembom et al. (2007) showed targeted likelihood estimates of variable importance measures could enhance biomarker discovery procedures. We have here introduced similar locally efficient doubly robust estimators suitable for right censored data structures, and also expect benefits to become apparent in real world applications.

References

[1] Bembom, B., Petersen, M.L., Rhee, S., Fessel, J.W., Sinisi, S.E., Shafer, R.W., and van der Laan, M.J. (2007). Biomarker Discovery Using Targeted Maximum Likelihood Estimation: Application to the Treatment of Antiretroviral Resistant HIV Infection. *U.C. Berkeley Division of Biostatistics Working Paper Series*, Working Paper 221.

[2] Bickel, P.J., Klaassen, C.A.J., Ritov, Y., and Wellner, J.A (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, New York.

[3] Breslow, N.E. (1974). Covariance analysis of censored survival data. *Biometrics.* 30, 89-99.

[4] Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B,* 34, 187-220.

[5] Kaplan, E.L., and P. Meier. (1958). Nonparametric estimator from incomplete observations. *Journal of the American Statistical Association.* 53, 457-481.

[6] Moore, K.L. and van der Laan M.J. (2007). Covariate Adjustment in Randomized Trials with Binary Outcomes: Targeted Maximum Likelihood Estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series*, Working Paper 215.

[7] Robins, J.M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine,* 16, 285-319.

[8] Robins, J.M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. *AIDS Methodology - Methodological Issues.* Jewell, N., Dietz, K, and Farewell, W., eds. Birkhäuser, Boston, 297-331.

[9] Robins J.M. and Rotnitzky, A. (2005). Inverse probability weighted estimation in survival analysis. *Encyclopedia of Biostatistics*, Second Edition, Editors: Armitage, P. and Colton, T., Wiley & Sons, New York.

[10] van der Laan, M.J. and Robins, J.M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer-Verlag, New York.

[11] van der Laan, M.J. and Rubin, D.B. (2006). Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics,* Vol. 2, Iss. 1, Article 11.

9.2 *Application of Time-to-Event Methods in the Assessment of Safety in Clinical Trials*

The following article appears as it was originally published on the University of California, Berkeley Division of Biostatistics Working Paper Series website in 2009, http://www.bepress.com/ucbbiostat/paper248/.

It was later published in the book **Design, Summarization, Analysis & Interpretation of Clinical Trials with Time-to-Event Endpoints**, edited by Karl E. Peace for Chapman and Hall in 2009.

Application of Time-to-Event Methods in the Assessment of Safety in Clinical Trials

Kelly L. Moore and Mark van der Laan

Abstract

Since randomized controlled trials (RCT) are typically designed and powered for efficacy rather than safety, power is an important concern in the analysis of the effect of treatment on the occurrence of adverse events (AE). These outcomes are often time-to-event outcomes which will naturally be subject to right-censoring due to early patient withdrawals. In the analysis of the treatment effect on such an outcome, gains in efficiency, and thus power, can be achieved by exploiting covariate information. We apply the targeted maximum likelihood methodology to the estimation of treatment specific survival at a fixed end point for right-censored survival outcomes. This approach provides a method for covariate adjustment, that under no or uninformative censoring, does not require any additional parametric modeling assumptions, and, under informative censoring, is consistent under consistent estimation of the censoring mechanism or the conditional hazard for survival. Thus, the targeted maximum likelihood estimator has two important advantages over the Kaplan-Meier estimator: 1) It exploits covariates to improve efficiency, and 2) It is consistent in the presence of informative censoring. These properties are demonstrated through simulation studies. Extensions to the methodology are provided for non randomized post-market safety studies and also for the inclusion of time-dependent covariates.

1 Introduction

Safety analysis in randomized controlled trials (RCT) involves estimation of the treatment effect on the numerous adverse events (AE) that are collected in the study. RCT are typically designed and powered for efficacy rather than safety. Even when assessment of AE is a major objective of study, the trial size is generally not increased to improve likelihood of detecting AE (Friedman et al. (1998)). As a result, power is an important concern in the analysis of the effect of treatment on AE in RCT (Peace (1987)).

Typically in an RCT, crude incidences of each AE are reported at some fixed end point such as the end of study (Gait et al. (2000); Güttner et al. (2007); Liu et al. (2006)). These crude estimates often ignore missing observations that frequently occur in RCT due to early patient withdrawals (Menjoge (2003)). A review of published RCT in major medical journals found that that censored data are often inadequately accounted for in their statistical analyses (Wood et al. (2004)). A crude estimator that ignores censoring can be highly biased when the proportion of dropouts differs between treatment groups (see Gait et al. (2000) for examples).

The crude incidence is an important consideration in the evaluation of safety for very rare, severe or unexpected AE. Such AE require clinical evaluation for each case and are not the focus of this paper. Instead, we focus on those AE that are routinely collected in RCT and most often are not associated with a pre-specified hypothesis. These AE are typically reported as an observed rate with a confidence interval or p-value.

Patient reporting of AE occurrence usually occurs at many intervals throughout the study often collected at follow-up interviews rather than only at a single fixed end-point. As such, time-to-event methods that exploit these data structures may provide further insight into the safety profile of the drug. The importance of considering estimators of AE rates that account for time due to differential lengths of exposure and follow-up is discussed in (O'Neill (1988)). Furthermore, in most RCT in oncology, most if not all patients suffer from some AE (Nishikawa et al. (2006)) and thus investigators may be interested in the probability of the occurrence of a given AE by a certain time rather than simply the incidence. Time-to-event analysis techniques may be more sensitive than crude estimates in that they readily handle missing observations that frequently occur in RCT due to early patient withdrawals. For example, in Davis et al. (1987), AE from the Beta-Blocker Heart Attack Trial were analyzed by comparing distributions of the time to the first AE in the two treatment arms. The results of this analysis were contrasted to the cross-sectional crude percentage analysis and were found to be more sensitive in detecting a difference by taking into account the withdrawals. A vast amount of literature exists for time-to-event analysis but these methods are often not applied to the analysis of AE in RCT. A general review of survival analysis methods in RCT (without a particular focus on AE) is provided in Fleming and Lin (2000).

In this paper we focus on estimation of treatment specific survival at a fixed end point for right-censored survival outcomes using targeted maximum likelihood estimation (van der Laan and Rubin (2006)). Survival is estimated based on a hazard fit and thus the time-dependent nature of the data is exploited. There are two main goals of the methodology presented in this paper over unadjusted crude proportions and Kaplan-Meier estimators. The first is to provide an estimator that exploits covariates to improve efficiency in the

estimation of treatment-specific survival at fixed end points. The second is to provide a consistent estimator in the presence of informative censoring.

2 Motivation and Outline

Consider the estimation of the effect of treatment on a particular AE at some fixed end point in the study. From estimation theory, it is known that the nonparametric maximum likelihood estimator (MLE) is the efficient estimator of the effect of interest (van der Laan and Robins (2003)). In most RCT, data are collected on baseline (pre-treatment) covariates in addition to the treatment and the AE of interest. The unadjusted or crude estimator is defined as the difference in proportions of the AE between treatment groups. This estimator ignores the covariates and is thus not equivalent to the full MLE. It follows that application of the unadjusted estimator can lead to a loss in estimation efficiency (precision) in practice.

Conflicting results in initial applications of covariate adjustment in RCT for estimating the treatment effect for fixed end-point efficacy studies were found. For continuous outcomes using linear models for adjustment demonstrated gains in precision over the unadjusted estimate (Pocock et al. (2002)). However adjustment using logistic models for binary outcomes was shown to actually reduce precision and inflate point estimates (Hernández et al. (2004); Robinson and Jewell (1991)).

This apparent contradiction was resolved through the application of estimating function methodology (Tsiatis et al. (2008); Zhang et al. (2008)) and targeted maximum likelihood estimation (Moore and van der Laan (2009)). In these references, consistent estimators that do not require parametric modeling assumptions were provided and shown to be more efficient than the unadjusted estimator, even with binary outcomes. It just so happens that the coefficient for the treatment variable in a linear regression that contains no interactions with treatment coincides with the efficient estimating function estimator and thus the targeted maximum likelihood estimator. This fortunate property does not hold for the logistic regression setting, i.e., the exponentiated coefficient for treatment from the logistic regression model does not equal the unadjusted odds ratio. This conditional estimator does not correspond to the marginal estimator in general and in particular not in the binary case. The efficient estimate of the marginal (i.e., unconditional) effect obtained from the conditional regression is the weighted average of the conditional effect of treatment on the outcome given covariates according to the distribution of the covariates.

With this principle of developing covariate adjusted estimators that do not require parametric modeling assumptions for consistency in mind, in this paper we provide a method for covariate adjustment in RCT for the estimation of treatment specific survival at a fixed end point for right-censored survival outcomes. Thereby, we can estimate a comparison of survival between treatment groups at a fixed end point that is some function of the two treatment specific survival estimates. Examples of such parameters are provided in section 4 such as the marginal additive difference in survival at a fixed end point. Under no or uninformative censoring, the estimator provided in this paper does not require any additional parametric modeling assumptions. Under informative censoring, the estimator is consistent under consistent estimation of the censoring mechanism or the conditional hazard for survival.

It is important to note that the conditional hazard on which the estimate is based is not meant to infer information about subgroup (conditional) effects of treatment. By averaging over the covariates that have terms in the hazard model, we obtain a marginal or unconditional estimate. The methodology presented in this paper can be extended to the estimation of subgroup specific effects however we focus only on marginal (unconditional) treatment effects on survival at fixed end point(s).

We also note that the methodology can be extended to provide a competitor test to the ubiquitous log-rank test. Methods have been proposed for covariate adjustment to improve power over the logrank test (Hernández et al. (2006); Li (2001); Lu and Tsiatis (2008)). These are tests for an average effect of treatment over time. Our efficiency results are not in comparison to these methods but rather to the treatment-specific Kaplan-Meier estimate at that fixed end point.

In itself treatment specific survival at a fixed end point, and thereby the effect of treatment on survival at that end point can provide useful information about the given AE of interest. This is a very common measure to report (see Gait et al. (2000); Güttner et al. (2007); Liu et al. (2006); Menjoge (2003)), however most of the currently applied estimation approaches ignore covariates and censoring and do not usually exploit the time-dependent nature of the data.

We present our method of covariate adjustment under the framework of targeted maximum likelihood estimation originally introduced in van der Laan and Rubin (2006). Specifically, the paper is outlined as follows. We first begin with a brief introduction to targeted maximum likelihood estimation in section 3. We then outline the data, model and parameter(s) of interest in section 4. The application of targeted maximum likelihood estimation to our parameter of interest with its statistical properties and inference are presented in section 5. In section 6 we present a simulation study to demonstrate the efficiency gains of the proposed method over the current methods in an RCT under no censoring and uninformative censoring. Furthermore, under informative censoring we demonstrate the bias that arises with the standard estimator in contrast to the consistency of our proposed estimator. The targeted maximum likelihood estimator requires estimation of an initial conditional hazard. Methods for fitting this initial hazard as well as the censoring mechanism are provided in section 7. In section 8 we outline the inverse weighting assumption for the censoring mechanism. Alternative estimators and their properties are briefly outlined in section 9. AE data are multivariate in nature in that many AE are collected an analyzed in any given RCT. In section 10 we outline the multiple testing issues involved in the analysis of such data. Section 11 provides extensions to the methodology including time-dependent covariates, and post-market safety analysis. Finally, we conclude with a discussion in section 12.

3 Introduction to targeted maximum likelihood estimation

Traditional maximum likelihood estimation aims for a trade-off between bias and variance for the whole density of the observed data O, whereas investigators are typically interested in a specific parameter of the density of O rather than the whole density itself. In this sec-

tion we discuss the algorithm generally, for technical details about this estimation approach we refer the reader to its seminal article (van der Laan and Rubin (2006)).

Define a model $\mathcal{M}$ which is a collection of probability distributions of $O \sim p_0$ and let $\hat{p}$ be an initial estimator of p_0. We are interested in a particular parameter of the data, $\psi_0 = \psi(p_0)$. To estimate this parameter, the targeted maximum likelihood algorithm's goal is to find a density $\hat{p}^* \in \mathcal{M}$ that solves the efficient influence curve estimating equation for the parameter of interest that results in a bias reduction in comparison to the maximum likelihood estimate $\psi(\hat{p})$ but also to find $\hat{p}^*$ that increases the log-likelihood relative to $\hat{p}$.

To estimate this $\hat{p}^*$, the algorithm finds a fluctuation of the initial $\hat{p}$ that results in a maximum change in ψ by constructing a path denoted by $\hat{p}(\epsilon)$ through $\hat{p}$ where ϵ is a free parameter. The score of this path at $\epsilon = 0$ equals the efficient influence curve. The optimal fluctuation is obtained by maximizing the likelihood of the data over ϵ and applying this fluctuation to $\hat{p}$ to obtain $\hat{p}^1$. This is the first step of the targeted maximum likelihood algorithm and the process is iterated until the fluctuation is essentially zero. The final step of the algorithm gives the targeted maximum likelihood estimate $\hat{p}^*$ which solves the efficient influence curve estimating equation and thus the resulting substitution estimator $\psi(\hat{p}^*)$ inherits the desirable properties of the estimating function based methodology, namely local efficiency and double robustness (van der Laan and Robins (2003)). It is also completely based on the maximum likelihood principle, resulting in robust finite sample behavior.

Targeted MLEs not only share the optimal properties with estimating equation estimators, but they also overcome some of their drawbacks. Estimating equation methodology requires that the efficient influence curve can be represented as an estimating function in terms of a parameter of interest and nuisance parameters which is not required by the targeted maximum likelihood algorithm since it simply solves the efficient influence curve estimating equation in p itself. Estimating equation estimators require external estimation of the nuisance parameters, while in the targeted maximum likelihood estimation procedure the estimator of the parameter of interest and the nuisance parameters are compatible with a single density estimator. Finally, estimating equation methodology lacks a criterion for selecting among candidate solutions in situations where multiple solutions in the parameter of interest exist, where the targeted maximum likelihood estimation approach can use the likelihood criterion to select among the targeted MLEs indexed by initial density estimators.

4 Data, Model and Parameter of Interest

We assume that in the study protocol, each patient is monitored at K equally spaced clinical visits. At each visit, M AE are evaluated as having occurred or not occurred. We focus on the first occurrence of the AE and thus let T represent the first visit when the AE reported as occurring and thus can take values $\{1, ..., K\}$. The censoring time C is the first visit when the subject is no longer enrolled in the study. Let $A \in \{0, 1\}$ represent the treatment assignment at baseline and W represents a vector of baseline covariates. The observed data are given by $O = (\tilde{T}, \Delta, A, W) \sim p_0$ where $\tilde{T} = \min(T, C)$, $\Delta = I(T \leq C)$ is the indicator that that subject was not censored and p_0 denotes the density of O. The conditional hazard is given by $\lambda_0(\cdot \mid A, W)$ and the corresponding conditional survival is

given by $S_0(\cdot \mid A, W)$. We present the methodology for estimation of the treatment effect for a single AE out of the M total AE collected. This procedure would be repeated for each of the M AE. For multiplicity considerations see section 10.

Let T_1 represent a patient's time to the occurrence of an AE had she possibly contrary to fact been assigned to the treatment group and let T_0 likewise represent the time to the occurrence of the AE had the patient been assigned to the control group.

Let $\mathcal{M}$ be the class of all densities of O with respect to an appropriate dominating measure where $\mathcal{M}$ is nonparametric up to possible smoothness conditions. Let our parameter of interest be represented by $\Psi(p_0)$. Specifically, we aim to estimate the following treatment specific parameters,

$$P_0 \rightarrow \Psi_1(p_0)(t_k) = Pr(T_1 > t_k) = E_0(S_0(t_k|A = 1, W)), \tag{1}$$

and

$$P_0 \rightarrow \Psi_0(p_0)(t_k) = Pr(T_0 > t_k) = E_0(S_0(t_k|A = 0, W)), \tag{2}$$

where the subscript for Ψ denotes the treatment group, either 0 or 1. In order to estimate the effect of treatment A on survival T we can thereby estimate a parameter that is some combination of $Pr(T_1 > t_k)$ and $Pr(T_0 > t_k)$. Examples include the marginal log hazard of survival, the marginal additive difference in the probability of survival, and the marginal log relative risk of survival at a fixed time t_k given respectively by,

$$P_0 \rightarrow \Psi_{HZ}(p_0)(t_k) = \log\left(\frac{\log(Pr(T_1 > t_k))}{\log(Pr(T_0 > t_k))}\right), \tag{3}$$

$$P_0 \rightarrow \Psi_{AD}(p_0)(t_k) = Pr(T_1 > t_k) - Pr(T_0 > t_k), \tag{4}$$

and

$$P_0 \rightarrow \Psi_{RR}(p_0)(t_k) = \log\left(\frac{Pr(T_1 > t_k)}{Pr(T_0 > t_k)}\right). \tag{5}$$

We note that if one averaged $\Psi_{HZ}(p_0)(t_k)$ over t, this would correspond with the Cox proportional hazards parameter and thus the parameter tested by the log rank test. However, we focus only on the t_k-specific parameter in this paper.

5 Targeted maximum likelihood estimation of marginal treatment specific survival at a fixed end point

Consider an initial fit $\hat{p}^0$ of the density of the observed data O identified by a hazard fit $\hat{\lambda}^0(t \mid A, W)$, the distribution of A identified by $\hat{g}^0(1 \mid W)$ and $\hat{g}^0(0 \mid W) = 1 - \hat{g}^0(1 \mid W)$, the censoring mechanism $\hat{G}^0(t \mid A, W)$ and the marginal distribution of W being the empirical probability distribution of $W_1, ..., W_n$. In an RCT, treatment is randomized and $\hat{g}^0(1|W) = \frac{1}{n}\sum_{i=1}^{n} A_i$.

Let the survival time be discrete and let the initial hazard fit $\hat{\lambda}(t \mid A, W)$ be given by a logistic regression model,

$$\text{logit}(\hat{\lambda}(t \mid A, W)) = \hat{\alpha}(t) + m(A, W \mid \hat{\beta}),$$

where m is some function of A and W. The targeted maximum likelihood estimation algorithm updates this initial fit by adding to it the term $\epsilon h(t, A, W)$, i.e.,

$$\text{logit}(\hat{\lambda}(\epsilon)(t \mid A, W)) = \hat{\alpha}(t) + m(A, W|\hat{\beta}) + \epsilon h(t, A, W). \tag{6}$$

The algorithm selects $h(t, A, W)$ such the score for this hazard model at $\epsilon = 0$ is equal to the projection of the efficient influence curve on scores generated by the parameter $\lambda(t \mid A, W))$ in the nonparametric model for the observed data assuming only coarsening at random (CAR).

The general formula for this covariate $h(t, A, W)$ for updating an initial hazard fit was provided in van der Laan and Rubin (2007) and is given by,

$$h(t, A, W) = \frac{D^{FULL}(A, W, t \mid \hat{p}) - E_{\hat{p}}[D^{FULL}(A, W, T \mid \hat{p}) \mid A, W, T > t)]}{\bar{G}(t_- \mid A, W)}, \tag{7}$$

where D^{FULL} is the efficient influence curve of the parameter of interest in the model in which there is no right censoring. This is also the optimal estimating function in this model. This full data estimating function for $\Psi_1(p_0)(t_k)$ provided in equation 1 is given by,

$$D_1^{FULL}(T, A, W \mid p)(t_k) = [I(T > t_k) - S(t_k \mid A, W)]\frac{I(A = 1)}{g(1|W)} + S(t_k \mid 1, W) - \psi_1(p), \tag{8}$$

and for $\Psi_0(p_0)(t_k)$ provided in equation 2 it is given by,

$$D_0^{FULL}(T, A, W \mid p)(t_k) = [I(T > t_k) - S(t_k \mid A, W)]\frac{I(A = 0)}{g(0|W)} + S(t_k \mid 0, W) - \psi_0(p), \tag{9}$$

To obtain the specific covariates for targeting the parameters $\Psi_1(p_0)(t_k)$ and $\Psi_0(p_0)(t_k)$, the full data estimating functions provided in equations 8 and 9 at $t = t_k$ are substituted into equation 7. Evaluating these substitutions gives the covariates,

$$h_1(t, A, W) = -\frac{I(A = 1)}{g(1)\bar{G}(t_- \mid A, W)}\frac{S(t_k \mid A, W)}{S(t \mid A, W)}I(t \leq t_k), \tag{10}$$

and

$$h_0(t, A, W) = -\frac{I(A = 0)}{g(0)\bar{G}(t_- \mid A, W)}\frac{S(t_k \mid A, W)}{S(t \mid A, W)}I(t \leq t_k), \tag{11}$$

for the treatment specific parameters $\Psi_1(p_0)(t_k)$ and $\Psi_0(p_0)(t_k)$ respectively.

Finding $\hat{\epsilon}$ in the updated hazard provided in equation 6 to maximize the likelihood of the observed data can be done in practice by fitting a logistic regression in the covariates $m(A, W \mid \hat{\beta})$ and $h(t, A, W)$. The coefficient for $m(A, W \mid \hat{\beta})$ is fixed at one and the intercept is set to zero and thus the whole regression is not refit, rather only ϵ is estimated.

These steps for evaluating $\hat{\epsilon}$ correspond with a single iteration of the targeted maximum likelihood algorithm. In the second iteration, the updated $\hat{\lambda}^1(t \mid A, W)$ now plays the role of the initial fit and the covariate $h(t, A, W)$ is then re-evaluated with the updated $\hat{S}^1(t \mid A, W)$ based on $\hat{\lambda}^1(t \mid A, W)$. In the third iteration $\hat{\lambda}^2(t|A, W)$ is fit and the procedure is iterated until $\hat{\epsilon}$ is essentially zero. The final hazard fit at the last iteration of the algorithm is denoted by $\hat{\lambda}^*(t \mid A, W)$ with the corresponding survival fit given by $\hat{S}^*(t \mid A, W)$.

As we are estimating two treatment specific parameters, we could either carry out the iterative updating procedure for each parameter separately or update the hazard fit simultaneously. To update the fit simultaneously, both covariates are added to the initial fit, i.e.,

$$\text{logit}(\hat{\lambda}(\epsilon)(t \mid A, W)) = \hat{\alpha}(t) + m(A, W|\hat{\beta}) + \epsilon_1 h_1(t, A, W) + \epsilon_2 h_0(t, A, W).$$

The iterative procedure is applied by now estimating two coefficients in each iteration as described above until both ϵ_1 and ϵ_2 are essentially zero.

Finally, the targeted maximum likelihood estimates of the probability of surviving past time t_k for subjects in treatment arms 1 and 0 given by $\Psi_1(p_0)(t_k)$ and $\Psi_0(p_0)(t_k)$ are computed by,

$$\hat{\psi}_1^*(t_k) = \frac{1}{n} \sum_{i=1}^{n} \hat{S}^*(t_k \mid 1, W_i).$$

and

$$\hat{\psi}_0^*(t_k) = \frac{1}{n} \sum_{i=1}^{n} \hat{S}^*(t_k \mid 0, W_i).$$

5.1 Rationale for updating only initial hazard

The initial fit $\hat{p}^0$ of p_0 is identified by $\hat{\lambda}^0(t \mid A, W)$, $\hat{g}^0(A \mid W)$, $\hat{G}^0(t \mid A, W)$ and the marginal distribution of W. However the algorithm only updates $\hat{\lambda}^0(t \mid A, W)$. Assuming CAR the density of the observed data p factorizes in to the marginal distribution of W given by p_W, the treatment mechanism $g(A \mid W)$, the conditional probability of censoring up to time t given by $\bar{G}(t \mid A, W)$ and the product over time of the conditional hazard at $T = t$ given by $\lambda(t \mid A, W)$. This factorization implies the orthogonal decomposition of functions of O in the Hilbert space $L^2(p)$. We can thus apply this decomposition to the efficient influence curve $D(O \mid p)$. As shown in van der Laan and Robins (2003), $D(O \mid p)$ is orthogonal to the tangent space $T_{CAR}(p)$ of the censoring and treatment mechanisms. Thus the components corresponding with $g(A \mid W)$ and $\bar{G}(t \mid A, W)$ are zero. This leaves the non zero components p_W and $\lambda(t \mid A, W)$. We choose the initial empirical distribution for W to estimate p_W which is the nonparametric maximum likelihood estimate for p_W and is therefore not updated. Thus the only element that does require updating is $\hat{\lambda}^0(t \mid A, W)$.

The efficient influence curve for $\Psi_1(p_0)(t_k)$ can be represented as,

$$\begin{aligned} D_1(p_0) \;=\; & \sum_{t<=t_k} h_1(g_0, G_0, S_0)(t, A, W)[I(\tilde{T} = t, \Delta = 1) - I(\tilde{T} >= t)\lambda_0(t \mid A = 1, W)] \\ & + S_0(t_k \mid A = 1, W) - \Psi_1(p_0)(t_k), \end{aligned} \tag{12}$$

where $S_0(t_k \mid A = 1, W)$ is a transformation of $\lambda_0(t \mid A = 1, W)$. This representation demonstrates the orthogonal decomposition described above. The empirical mean of the second component of $D_1(p_0)$ given by $S_0(t_k \mid A = 1, W) - E_0 S_0(t_k \mid A = 1, W)$ is always solved by using empirical distribution to estimate the marginal distribution of W. Thus the targeted maximum likelihood estimator solves this second component. The first component, the covariate times the residuals, is solved by performing the iterative targeted maximum likelihood algorithm with logistic regression fit of the discrete hazard $\lambda_0(t \mid A, W)$. We note that similarly, the efficient influence curve for $\Psi_0(p_0)(t_k)$ can be represented as,

$$\begin{aligned} D_0(p_0) &= \sum_{t<=t_k} h_0(g_0, G_0, S_0)(t \mid A, W)[I(\tilde{T} = t, \Delta = 1) - I(\tilde{T} >= t)\lambda_0(t \mid A = 0, W)] \\ &\quad + S_0(t_k \mid A = 0, W) - \Psi_0(p_0)(t_k). \end{aligned} \tag{13}$$

5.2 Statistical Properties

The targeted maximum likelihood estimate $\hat{p}^* \in \mathcal{M}$ of p_0 solves the efficient influence curve which is the optimal estimating equation for the parameter of interest. It can be shown that $E_0 D_1(p_0) = E_0 D_1(S, g, G) = 0$ if either $S = S(\cdot \mid A, W)$ (and thus $\lambda(\cdot \mid A, W)$) is consistently estimated or $g_0(A \mid W)$ and $\bar{G}_0(\cdot \mid A, W)$ are consistently estimated. When the treatment is assigned completely at random as in an RCT, the treatment mechanism is known and $g(A \mid W) = g(A)$. Thus consistency of $\hat{\psi}_1^*(t_k)$ in an RCT relies on only consistent estimation of $\bar{G}_0(\cdot \mid A, W)$ or $S(\cdot \mid A, W)$. When there is no censoring or censoring is missing completely at random (MCAR), $\hat{\psi}_1^*(t_k)$ is consistent even when the estimator $\hat{S}(\cdot \mid A, W)$ of $S(\cdot \mid A, W)$ is inconsistent (e.g., if it relies on a misspecified model). One is hence not concerned with estimation bias with this method in an RCT. Under informative or missing at random (MAR) censoring, if $\bar{G}_0(\cdot \mid A, W)$ is consistently estimated then $\hat{\psi}_1^*(t_k)$ is consistent even if $\hat{S}(\cdot \mid A, W)$ is mis-specified. If both are correctly specified then $\hat{\psi}_1^*(t_k)$ is efficient. These same statistical properties also hold for $\hat{\psi}_0^*(t_k)$.

5.3 Inference

Let $\hat{p}^*$ represent the targeted maximum likelihood estimate of p_0. One can construct a Wald-type 0.95-confidence interval for $\hat{\psi}_1^*(t_k)$ based on the estimate of the efficient influence curve $D_1(\hat{p}^*)(O)$ where $D_1(p)$ is given by equation 12. The asymptotic variance of $\sqrt{n}(\hat{\psi}_1^*(t_k) - \Psi_1(p_0)(t_k))$ can be estimated with

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} D_1^2(\hat{p}^*)(O_i).$$

The corresponding asymptotically conservative Wald-type 0.95-confidence interval is defined as $\hat{\psi}_1^*(t_k) \pm 1.96 \frac{\hat{\sigma}}{\sqrt{n}}$. The null hypothesis $H_0 : \Psi_1(p_0)(t_k) = 0$ can be tested with the test statistic

$$T_n = \frac{\hat{\psi}_1^*(t_k)}{\frac{\hat{\sigma}}{\sqrt{n}}},$$

whose asymptotic distribution is $N(0,1)$ under the null hypothesis. Similarly, confidence intervals and test statistics for $\Psi_0(p_0)(t_k)$ can be computed based on the estimate of the efficient influence curve $D_0(\hat{p}^*)(O)$ where $D_0(p)$ is given by equation 13.

If our parameter of interest is some function of the treatment specific survival estimates we can apply the δ-method to obtain the estimate of its influence curve. Specifically the estimated influence curve for the log hazard of survival, additive difference in survival, and relative risk of survival at t_k provided in equations 3, 4, and 5 are respectively given by,

1. $\Psi_{HZ}(p_0)(t_k) : \frac{1}{\hat{\psi}_1^*(t_k)\log(\hat{\psi}_1^*(t_k))} D_1(\hat{p}^*)(O) - \frac{1}{\hat{\psi}_0^*(t_k)\log(\hat{\psi}_0^*(t_k))} D_0(\hat{p}^*)(O)$
2. $\Psi_{AD}(p_0)(t_k) : D_1(\hat{p}^*)(O) - D_0(\hat{p}^*)(O)$
3. $\Psi_{RR}(p_0)(t_k) : -\frac{1}{1-\hat{\psi}_1^*(t_k)} D_1(\hat{p}^*)(O) + \frac{1}{1-\hat{\psi}_0^*(t_k)} D_0(\hat{p}^*)(O)$

We can again compute confidence intervals and test statistics for these parameters using the estimated influence curve to estimate the asymptotic variance.

As an alternative to the influence curve based estimates of the asymptotic variance, one can obtain valid inference using the bootstrap procedure.

The inference provided in this section is for the estimates of the treatment effect for a single AE. For multiplicity adjustments for the analysis of a set of AE see section 10.

6 Simulation Study

The targeted maximum likelihood estimation procedure was applied to simulated data to illustrate the estimator's potential gains in efficiency. The conditions under which the greatest gains can be achieved over the standard unadjusted estimator were explored in addition to the estimators' performance in the presence of informative censoring.

6.1 Simulation Protocol

We simulated 1000 replicates of sample size 300 from the following data generating distribution where time is discrete and takes values $t_k \in \{1, ..., 10\}$:

- $Pr(A = 1) = Pr(A = 0) = 0.5$
- $W \sim U(0.2, 1.2)$
- $\lambda(t|A, W) = \frac{I(t_k<10)I(Y(t_k-1)=0)}{1+\exp(-(-3-A+\beta_W W^2))} + I(t_k = 10)$
- $\lambda_C(t|A, W) = \frac{I(\Delta(t_k-1)=0)}{1+\exp(-(-\gamma_0-\gamma_1 A-\gamma_2 W))}$,

where $\lambda(t|A, W)$ is the hazard for survival and $\lambda_C(t|A, W)$ is the hazard for censoring. Two different data generating hazards for survival were applied corresponding with two values for β_W. These two values were set to $\beta_W \in \{1, 3\}$ corresponding with correlations between W and failure time of -0.22 and -0.63 respectively. We refer to the simulated data with $\beta_W = 1$ as the weak covariate setting and $\beta_W = 3$ as the strong covariate setting.

Three different types of censoring were simulated, no censoring, MCAR and MAR. Each type of censoring was applied to the weak and strong covariate settings for a total of six simulation scenarios. For both the weak and strong covariate settings, the MCAR an MAR censoring mechanisms were set such that approximately 33% of the observations were censored. The censoring was generated to ensure that $\bar{G}(t|A,W) > 0$ (see section 8 for details of this assumption). If censoring and failure time were tied, the subject was considered uncensored. For a summary of the simulation settings and the specific parameter values, see Table 1.

Table 1: Summary of simulation settings. "Corr" is correlation, $\gamma = (\gamma_0, \gamma_1, \gamma_2)$ are the coefficients for the hazard for censoring, and β_W is the coefficient for W in the hazard for survival.

Scenario	Censoring	γ	Corr W and T	β_W
1	No censoring	NA	-0.22 (Weak)	1
2	MCAR	(-2.7,0,0)	-0.22 (Weak)	1
3	MAR	(-1.65,0.5,-2)	-0.22 (Weak)	1
4	No censoring	NA	-0.65 (Strong)	3
5	MCAR	(-2,0,0)	-0.65 (Strong)	3
6	MAR	(-1.15,0.5,-2)	-0.65 (Strong)	3

The difference in treatment-specific survival probabilities given by $\psi(t_k) = E_0(S_0(t_k|A = 1, W) - S_0(t_k|A = 0, W))$ was estimated at each time point $t_k = 1$ through $t_k = 9$. The unadjusted estimator is defined as the difference in the treatment specific Kaplan-Meier estimators at t_k. The targeted maximum likelihood estimator was applied using two different initial hazard fits. The first initial hazard was correctly specified. The second initial hazard was mis-specified by including A and W as main terms and an interaction term between A and W. For both initial hazard fits, only time points 1 through 9 were included in the fit as the AE had occurred for all subjects by time point 10 and thus the hazard was one at $t_k = 10$. In the MCAR censoring setting, the censoring mechanism was estimated using Kaplan-Meier. In the MAR censoring setting, the censoring mechanism was correctly specified. The update of the initial hazard was performed by adding to it the two covariates h_1 and h_0 provided in equations 10 and 11 respectively. The corresponding coefficients ϵ_1 and ϵ_2 were simultaneously estimated by fixing the offset from the initial fit and setting the intercept to 0. The procedure was iterated until ϵ_1 and ϵ_2 were sufficiently close to zero.

The estimators were compared using a relative efficiency measure based on the mean squared error (MSE) computed as the MSE of the unadjusted estimates divided by the MSE of the targeted maximum likelihood estimates. Thus a value greater than one indicates a gain in efficiency of the covariate adjusted targeted maximum likelihood estimator over the unadjusted estimator.

In addition to these six simulation scenarios, to explore the relationship between relative efficiency and the correlation between the covariate and failure time, we generated data by varying β_W in the data generating distribution above for six values, $\beta_W \in \{0.5, 1, 1.5, 2, 2.5, 3\}$ corresponding with correlations between W and failure time of {-0.10,-

0.22,-0.36,-0.46,-0.56,-0.63} under no censoring. The parameter $\psi(5)$ was estimated based on 1000 sampled datasets with sample size $n = 300$.

6.2 Simulation Results and Discussion

6.2.1 Strong covariate setting

In the no censoring and MCAR censoring scenarios, the bias should be approximately zero. Thus, the relative MSE is essentially comparing the variance of the unadjusted and targeted maximum likelihood estimates. Any gain in the MSE can therefore be attributed to a reduction in variance due to the covariate adjustment. In this strong covariate setting, exploiting this covariate by applying the targeted maximum likelihood estimator should provide a gain precision due to a reduction in the residuals. In the informative censoring setting (MAR), in addition to the expected gain in efficiency we expect a reduction in bias of the targeted maximum likelihood estimator with the correctly specified treatment mechanism over the unadjusted estimator. The informative censoring is accounted for through the covariates h_1 and h_0 that are inverse weighted by the subjects' conditional probability of being observed at time t given their observed history.

Figure 1 provides the relative MSE results for $\hat{\psi}(t_k)$ for $t_k \in \{1, ...9\}$ for the strong covariate setting with $\beta_W = 3$. Based on these results, we observe that indeed the expected gain in efficiency is achieved. The minimum observed relative MSE was 1.25 for $t_k = 1$ in the MAR censoring setting with a mis-specified initial hazard fit. A maximum relative MSE of 1.9 is observed under the no censoring setting with the correctly specified initial hazard at $t_k = 3$. The approximate overall average relative MSE was 1.6. Consistently across all time points and censoring scenarios, the targeted maximum likelihood estimator is outperforming the unadjusted estimator.

Figure 2 provides the bias as a percent of the truth for the two estimators under the MAR censoring setting with the correctly specified initial hazard. Clearly as t_k increases, the bias of the unadjusted estimates increases whereas the targeted maximum likelihood estimates is relatively close to zero in comparison. Thus the targeted maximum likelihood approach can not only provide gains in efficiency through covariate adjustment, but can also account for informative censoring as well.

6.2.2 Weak covariate setting

In this weak covariate setting, again in the no censoring and MCAR censoring scenarios, the bias should essentially be zero. However, we expect a lesser gain in efficiency if any as compared to the strong covariate setting since the covariate in this setting is not as useful for hazard prediction. We do again expect a bias reduction in the MAR censoring setting for the targeted maximum likelihood estimator over the unadjusted estimator.

Figure 3 provides the relative MSE results for the weak correlation simulation with $\beta_W = 1$. As expected, the relative MSE are all close to one indicating that only small efficiency gains are achieved when only weak covariates are present in the data. However, as small the gains are they are also achieved across all time points as in the strong covariate setting. Regardless of the correlation between the covariate and failure time, in the informative

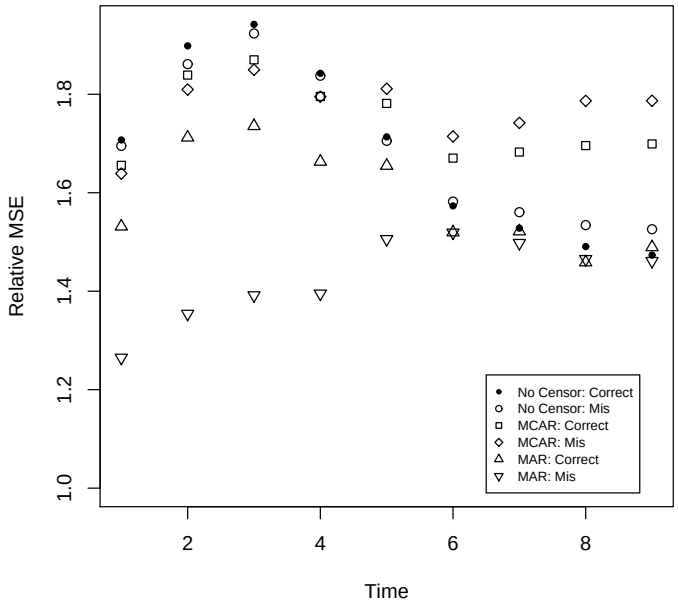


Figure 1: Relative MSE: Strong covariate setting ($\beta_W = 3$)

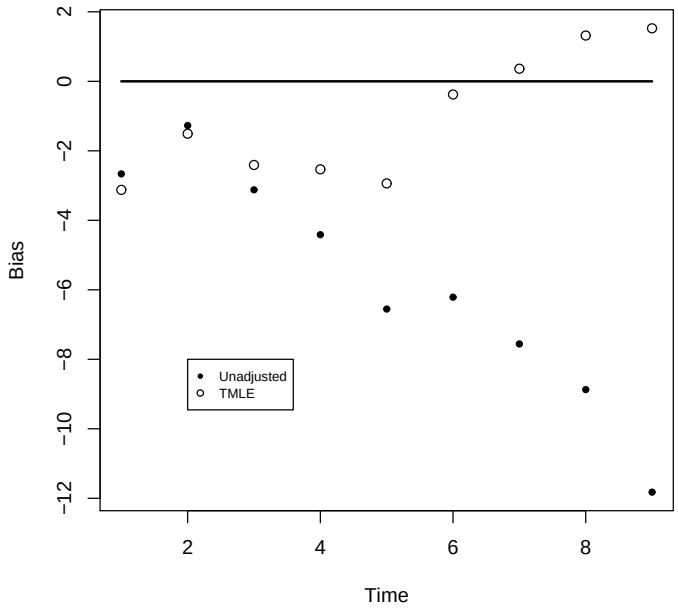


Figure 2: Bias: Strong covariate setting ($\beta_W = 3$) with informative censoring

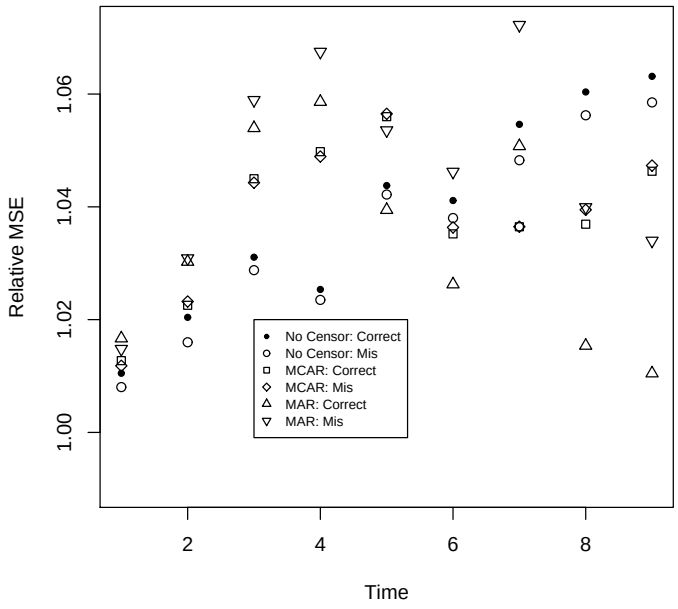


Figure 3: Relative MSE: Weak covariate setting ($\beta_W = 1$)

censoring scenario the targeted maximum likelihood estimate is consistent under consistent estimation of the censoring mechanism as evidenced in the plot of the % bias in Figure 4.

6.2.3 Relationship between correlation of covariate(s) and failure time with efficiency gain

As the correlation between W and failure time increases we expect to observe increasing gains in efficiency. Selecting an arbitrarily selected time point $t_k = 5$ for ease of presentation, Figure 5 clearly demonstrates that as the correlation between W and failure time increases so does the relative MSE. In fact in for this particular data generating distribution, at time $t_k = 5$ the relationship is nearly linear. These results reflect similar findings in RCT with fixed-end point studies where relations between R^2 and efficiency gain have been demonstrated (Moore and van der Laan (2009); Pocock et al. (2002)). This relationship indicates that if indeed the particular dataset contains covariates that are predictive of the failure time of the AE of interest, one can achieve gains in precision and thus power by using the targeted maximum likelihood estimator.

7 Fitting initial hazard and censoring mechanism

Despite these potential gains in efficiency as demonstrated by theory and simulation results, there has been concern with covariate adjustment in RCT with respect to investigators se-

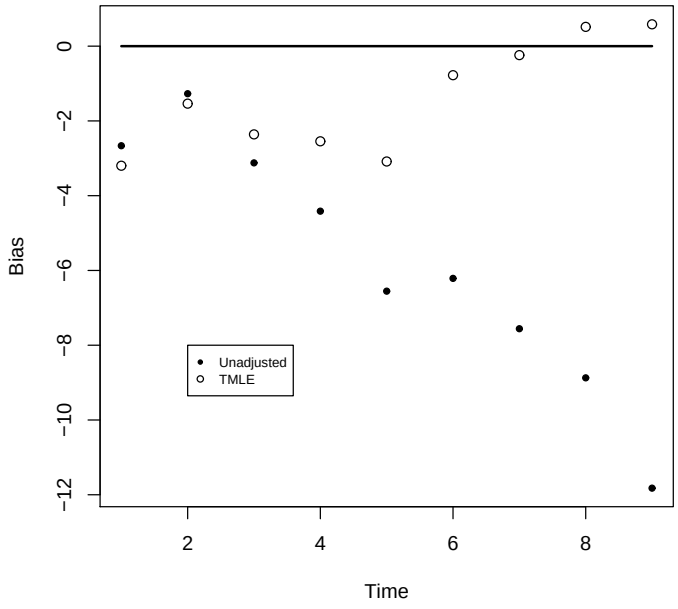


Figure 4: Bias: Weak covariate setting ($\beta_W = 1$) with informative censoring

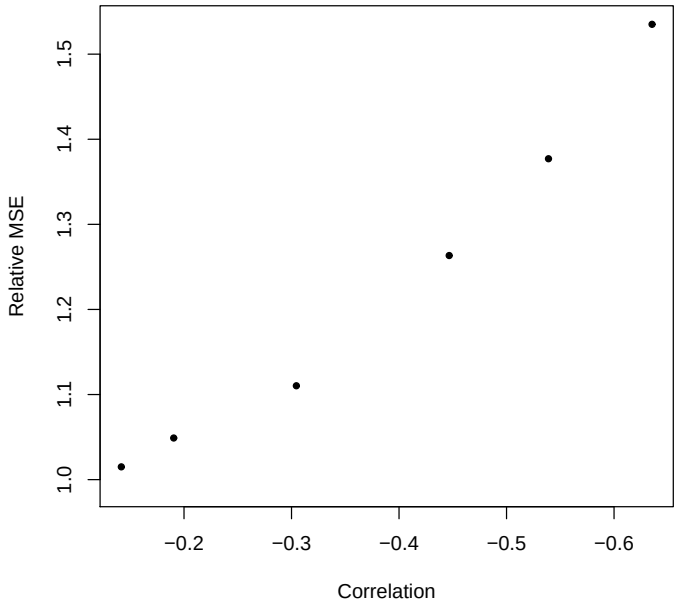


Figure 5: Efficiency gain and correlation between covariate and failure time

lecting covariates to obtain favorable inference. We conjecture that such cheating can be avoided if one uses an a priori specified algorithm for model selection. When the model selection procedure is specified in an analysis protocol, the analysis is protected from investigators guiding causal inferences based on selection of favorable covariates and their functional forms in a parametric model. In safety analysis, if investigator (sponsor) bias does indeed exist, it would be reasonable to assume that it would lean towards the treatment having no effect on the AE and thus the concerns are the reverse from efficacy analysis. The investigator bias would tend towards the less efficient unadjusted estimator. The analysis of AEs is often exploratory in nature and the results are meant to flag potential AE of concern which may reduce the motivation for dishonest inference using covariate adjustment. Regardless of the covariate selection strategy, it should be explicitly outlined to avoid any such concerns.

There are a number of model selection algorithms that can be applied to data-adaptively select the initial hazard fit. One such approach is the D/S/A algorithm that searches through a large space of functional forms using deletion, substitution and addition moves. One can apply this algorithm to the pooled data (over time) to fit the initial hazard (Sinisi and van der Laan (2004)). One can also fit hazards using the hazard regression (HARE) algorithm developed by Kooperberg et al. (1995), which uses piecewise linear regression splines and adaptively selects the covariates an knots. As another alternative, one could also include all covariates that have a strong univariate association with failure time in a hazard fit as main terms in addition to the treatment variable. Since one is often investigating many AE, a fast algorithm such as the latter may be an appropriate alternative for computational efficiency.

We also note that if weights are required as they are for the inverse probability of censoring weighted (IPCW) reduced data targeted maximum likelihood estimators as outlined in section 11.1, the D/S/A algorithm can be run with the corresponding weights.

In addition to the hazard for survival, the hazard for censoring must also be estimated. One of the algorithms discussed above can also be applied to estimate the censoring mechanism. We note that the application of the targeted maximum likelihood estimator to a set of M AE requires M hazard fits whereas only one fit for censoring is required. Thus, the censoring mechanism is estimated once and for all and is used in the analysis of each of the M AE.

8 Inverse weighting assumption

The targeted maximum likelihood estimator, as well as other inverse weighted estimators (see section 9) for the parameters presented in this paper rely on the assumption that each subject has a positive probability of being observed (i.e., not censored) at time t, which can be expressed by,

$$\bar{G}(t_- \mid A, W) > 0, t = t_k.$$

This identifiability assumption has been addressed as an important assumption for right-censored data (Robins and Rotnitzky (1992)). In Neugebauer and van der Laan (2005) it was demonstrated that practical violations of this assumption can result in severely variable and biased estimates.

One is alerted of such violations by observing very small probabilities of remaining uncensored based on the estimated censoring mechanism, i.e., there are patients with a probability of censoring of almost one given their observed past.

9 Alternative Estimators

Prior to the introduction of targeted maximum likelihood estimation, there were two main approaches to estimating the treatment specific survival at a fixed end point t_k: maximum likelihood estimation and estimating function estimation. In the maximum likelihood approach, one obtains an estimate $\hat{p}$ for p identified by perhaps a Cox proportional hazards model for continuous survival or logistic regression for discrete survival. The parameter of interest is then evaluated via substitution, i.e., $\hat{\psi} = \psi(\hat{p})$. These maximum likelihood substitution estimators involve estimating some hazard fit using an *a priori* specified model or a model selection algorithm that is concerned with performing well with respect to the whole density rather than the actual parameter of interest, e.g., the difference in treatment specific survival at a specific time t_k. These type of estimators often have poor performance and can be heavily biased whenever the estimated hazard is inconsistent (Robins and Ritov (1997)). Furthermore, inference for such maximum likelihood estimators that rely on parametric models are overly optimistic and thus their corresponding p-values are particularly unreliable. This is in contrast to the inference for the targeted maximum likelihood estimators which respects that no *a priori* models are required.

An alternative to the likelihood based approach is the extensively studied estimating function based approach. Recall that the full data estimating functions provided in equations 8 and 9 are estimating functions that could be applied to estimate the treatment specific survival at time t_k if we had access to the full data, i.e., the uncensored survival time. The full data estimating function can be mapped into a an estimating function based on the observed data using the IPCW method. The IPCW estimators based on the IPCW estimating function denoted by $D^{IPCW}(T, A, W \mid \psi_1, g, G)$ have been shown to be consistent and asymptotically linear if the censoring mechanism G can be well approximated (Robins and Rotnitzky (2005); van der Laan and Robins (2003)). While the IPCW estimators have advantages such as simple implementation, they are not optimal in terms of robustness and efficiency. Their consistency relies on correct estimation of the censoring mechanism whereas maximum likelihood estimators rely on correct estimation of the full likelihood of the data.

The efficient influence curve can be obtained by subtracting from the IPCW estimation function the IPCW projection onto the tangent space T_{CAR} of scores of the nuisance parameter G (van der Laan and Robins (2003)). The efficient influence curve is the optimal estimating function in terms of efficiency and robustness and the corresponding solution to this equation is the so-called double robust IPCW (DR-IPCW) estimator. The "double" robust properties of this estimator are equivalent to those of the targeted maximum likelihood estimator as the targeted maximum likelihood estimator solves the efficient influence curve estimating equation, see section 5.2. Despite the advantageous properties of such efficient estimating function based estimators, maximum likelihood based estimators are much more common in practice.

The more recently introduced targeted maximum likelihood estimation methodology that was applied in this paper can be viewed as a fusion between the likelihood and estimating function based methods. A notable advantage of the targeted maximum likelihood estimators is their relative ease of implementation in comparison to estimating equations which are often difficult to solve.

10 Multiple Testing considerations

An important consideration in safety analysis is multiple testing in that often as many as hundreds of AE are collected. The ICH guidelines indicate that it is recommended to adjust for multiplicity when hypothesis tests are applied (ICH (1996)). However, the ICH guidelines do not provide any specific methods for adjustment. The need for adjustment is demonstrated by the following example outlined in Kaplan et al. (2002). In this study, out of 92 safety comparisons the investigators found a single significant result according to unadjusted p-values. A larger hypothesis driven study for this AE that had no known clinical explanation was carried out and did not result in any significant findings. Such false positive results for testing the effect of treatment on a series of AE based on unadjusted p-values can cause undue concern for approval/labeling and can affect post-marketing commitments. On the other hand, over adjusting could also result in missing potentially relevant AE. Thus appropriate adjustment requires some balance between no adjustment and a highly stringent procedure such as Bonferroni.

Many advances have been made in the area of multiple testing over the Bonferroni-type methods including resampling based methods to control the familywise error rate (FWER), for example see van der Laan et al. (2004) and the Benjamini-Hochberg method for controlling the false discovery rate (FDR) (Benjamini and Hochberg (1995)). With FWER approaches, one is concerned with controlling the probability of erroneously rejecting one or more of the true null hypotheses, whereas the FDR approach controls the expected proportion of erroneous rejections among all rejections. The resampling based FWER method makes use of the correlation of test statistics which can provide a gain in power over assuming independence. However, the Benjamini-Hochberg FDR approach has been shown to be perform well with correlated test statistics as well (Benjamini et al. (1997)). The selection of the appropriate adjustment depends on whether or not a more conservative approach is reasonable. In safety analysis, one certainly does not want to miss flagging an important AE and thus might lean towards an FDR approach.

FDR methods have been proposed specifically in the analysis of AE in Mehrotra and Heyse (2004). Their method involves a two-step procedure that groups AE by body system and performs an FDR adjustment both within and across the body system. Presumably this method attempts to account for the dependency of the AE by grouping in this manner. Thus the multiple testing considerations and the dependency of the test statistics in safety analysis has indeed received some attention in literature.

The multiple testing adjustment procedure to be applied in the safety analysis should be provided in the study protocol to avoid potential for dishonest inference. In addition, the unadjusted p-values should continue to be reported with the adjusted p-values so all AE can be evaluated to assess their potential clinical relevance.

11 Extensions

11.1 Time-dependent covariates

It is not unlikely that many time-dependent measurements are collected at each follow-up visit in addition to the many AE and efficacy outcome measurements. Such time-dependent covariates are often predictive of censoring. The efficiency and robustness results presented in this paper have been based on data structures with baseline covariates only. The targeted maximum likelihood estimation procedure for data structures with time-dependent covariates is more complex as demonstrated in van der Laan (2008). To overcome this issue and avoid modeling the full likelihood, van der Laan (2008) introduced IPCW reduced data targeted maximum likelihood estimators. We provide only an informal description of this procedure here, for details we refer readers to the formal presentation provided in van der Laan (2008).

In this framework, the targeted maximum likelihood estimation procedure is carried out for a reduced data structure X^r, which in this case is the data structure that only includes baseline covariates. The IPCW reduced data procedure differs from the procedure where X^r is the full data in that the log-likelihoods are weighted by a time-dependent stabilizing weight given by,

$$sw(t) = \frac{I(C > t)\bar{G}^r(t \mid X^r)}{\bar{G}(t \mid X)}.$$

This stabilizing weight is based on $\bar{G}^r(t \mid X^r)$ which is the censoring mechanism based on the reduced data structure that includes baseline covariates only and $\bar{G}(t \mid X)$ which is the censoring mechanism based on the complete data structure that includes time-dependent covariates.

In practice in estimation of the parameter $\psi(t_k) = E_0(S_0(t_k|A = 1, W) - S_0(t_k|A = 0, W))$, one must apply these weights anytime maximum likelihood estimation is performed. Thus, the IPCW reduced data targeted maximum likelihood estimation procedure differs from the standard targeted maximum likelihood procedure provided in section 5 in that each time the conditional hazard is fit it is weighted by $sw(t)$. These weights are time-specific and thus each subject receives a different weight at each point in time. The initial hazard estimate $\hat{\lambda}^0(t \mid A, W)$ is weighted by $sw(t)$. The algorithm then updates $\hat{\lambda}^0(t \mid A, W)$ by adding the time-dependent covariates $h_1(t, A, W$ and $h_0(t, A, W)$ and estimating their corresponding coefficients ϵ_1 and ϵ_2. In the IPCW reduced data targeted maximum likelihood estimation procedure one includes the weights $sw(t)$ in estimation of ϵ_1 and ϵ_2. These weights are applied in each iteration of the algorithm to obtain the final fit $\hat{\lambda}^*(t \mid A, W)$ that is achieved when $\hat{\epsilon}_1$ and $\hat{\epsilon}_2$ are sufficiently close to zero. Thus estimation can again be achieved using standard software with the only additional requirement of weighting each of the regressions by these time-dependent weights.

Estimation of these time-dependent weights requires estimation of $\bar{G}^r(t \mid X)$ and $\bar{G}(t \mid X)$. Model selection algorithms that can be applied to estimate $\bar{G}^r(t \mid X)$ were described in section 7. Similarly the censoring mechanism $\bar{G}(t \mid X)$ can be estimated using a Cox proportional hazards model with time-dependent covariates for continuous censoring times or logistic regression model with time dependent covariates for discrete censoring times. Model selection algorithms such as those described in section 7 can also be applied by

including these time-dependent covariates as candidates.

Let $\hat{\psi}^r(t_k)$ represent the IPCW reduced data targeted maximum likelihood estimator of $\psi(t_k)$. By applying this IPCW weighting in the reduced data targeted maximum likelihood estimation procedure a particular type of double robustness is obtained. If there are no time-dependent covariates that are predictive of censoring time, then the ratio of estimated survival probabilities of censoring in the above weight $sw(t)$ is one. In this case, if $\bar{G}(t \mid X)$ is consistently estimated or $\lambda(\cdot \mid A, W)$ is consistently estimated then $\hat{\psi}^r(t_k)$ is consistent; if both are consistent then it is even more efficient than the estimator that was based on the reduced data structure. If there are indeed time-dependent covariates that are predictive of censoring time, and $\bar{G}(t \mid A, W)$ is well approximated then $\hat{\psi}^r(t_k)$ is consistent and the desired bias reduction is achieved.

11.2 Post market data

As RCT are powered for efficacy, it is often the case that many AE are either not observed at all during the pre-market phase or so few are observed that statistically conclusive results are often exceptions (Peace (1987)). In an RCT of a rotavirus vaccine in which the AE of intussusception among vaccine recipients compared to controls was not found to be statistically significant. After the vaccine was approved and had been widely used, an association between this AE and the vaccine was found and it was pulled off the market. A subsequent analysis demonstrated that to obtain power of 50% to detect a difference as small as the actual observed Phase III incidence of the AE, a sample size of approximately 90,000 would be required (6 times the actual sample size) (Jacobson et al. (2001)). Due to the high cost and complications involved in running an RCT, such large sample sizes are not feasible.

It is not only the rarity of many AE that causes issues in detection during RCT, but also the fact that RCT may have restrictive inclusion criteria whereas the drug is likely applied to a less restrictive population in post-market. Furthermore, the follow-up time in the pre-market phase may not be long enough to detect delayed AE. For a discussion regarding the difficulties in "proving" safety of a compound in general see Bross (1985). Post-market monitoring is therefore an important aspect of safety analysis.

There are a number of types of post-market data (for a thorough description of the various types of post-market data see Glasser et al. (2007)) including spontaneous adverse event reporting systems (e.g., "MedWatch"). These data can be useful for detecting potentially new or unexpected adverse drug reactions that require further analysis however they often suffer from under-reporting by as much as a factor of 20 (Edlavitch (1988)).

In this section, we focus on observational post-market studies or pharmacoepidemiological studies. Since patients in these type of studies are not randomized to a drug versus placebo (or competitor), confounding is typically present. Of particular concern is the fact that sicker patients are often selected to receive one particular drug versus another. There exists a vast amount of literature for controlling for confounding in epidemiological studies. Popular methods in pharmacoepidemiology include propensity score (PS) methods and regression based approaches. However, consistency with these methods rely on correct specification of the PS or the regression model used. Furthermore, it is not clear how informative censoring is accounted for with these methods. The targeted maximum likelihood

estimators are double robust and are thus more advantageous than these commonly applied alternative approaches.

Before we proceed with discussion of estimation of causal effects with observational data, we first outline the data and assumptions. Suppose we observe n independent and identically distributed copies of $O = (\tilde{T}, \Delta, A, W) \sim p_0$ as defined in section 4. Causal effects are based on a hypothetical full data structure $X = (T_{1,1}, T_{1,0}, T_{0,1}, T_{0,0}, W)$ which is a collection of action specific survival times where this action is comprised of treatment and censoring. Note that we are only interested in the counterfactuals under this joint action-mechanism that consists of both censoring and treatment mechanisms where censoring equals zero, i.e., $T_{1,0}$ and $T_{0,0}$. In other words, we aim to investigate what would have happened under each treatment had censoring not occurred.

The consistency assumption states that the observed data consist of the counterfactual outcome corresponding with the joint action actually observed. The coarsening at random (CAR) assumption implies that the joint action is conditionally independent of the full data X given the observed data. We denote the conditional probability distribution of treatment A by $g_0(a \mid X) \equiv P(A = a \mid X)$. In observational studies, CAR implies $g_0(A \mid X) = g_0(A \mid W)$, in contrast to RCT in which treatment is assigned completely at random and $g_0(A \mid X) = g_0(A)$.

We aim to estimate $\psi(t_k) = E_0(S_0(t_k|A = 1, W) - S_0(t_k|A = 0, W)) = Pr(T_{1,0} > t_k) - Pr(T_{0,0} > t_k)$. Even under no censoring or MCAR, we are can no longer rely on the unadjusted treatment specific Kaplan-Meier estimates being unbiased due to confounding of treatment.

Under the assumptions above, the targeted maximum likelihood estimator for $\psi(t_k)$ is double robust and locally efficient. Thus the targeted maximum likelihood estimation procedure described in this paper is theoretically optimal in terms of robustness and efficiency. In our presentation, we assumed that treatment was assigned at random. In observational studies, in addition to estimating $\lambda(\cdot \mid A, W)$ and possibly $\bar{G}(\cdot \mid A, W)$ (when censoring is present), observational studies require estimation of the treatment mechanism $g(A \mid W)$ as well. It has been demonstrated that when censoring is MCAR in an RCT, the targeted maximum likelihood estimate $\hat{\psi}^*(t_k)$ is consistent under mis-specification of $\lambda(\cdot \mid A, W)$ since $g(A \mid W)$ is always correctly specified. However, even under MCAR, in observational studies, consistency of $\hat{\psi}^*(t_k)$ relies on consistent estimation of $\lambda(\cdot \mid A, W)$ or $g(A \mid W)$ and is efficient if both are consistently estimated (van der Laan and Rubin (2006)). When censoring is MAR, then consistency of $\hat{\psi}^*(t_k)$ also relies on consistent estimation of the joint missingness $g(A \mid W)$ and $\bar{G}(\cdot \mid A, W)$ or $\lambda(\cdot \mid A, W)$.

We also note that the targeted maximum likelihood estimators as well as the commonly applied PS methods rely on the experimental treatment assignment (ETA) assumption. Under this assumption, each patient must have a positive probability of receiving each treatment. The inverse weighted PS estimator is known to suffer severely from violations of this assumption in practice (Neugebauer and van der Laan (2005); Robins and Rotnitzky (1992); Wang et al. (2006)). This poor performance is evident with inverse weighting, however we note that all other PS methods rely on this assumption as well, but are not as sensitive to practical violations. This assumption is essentially about information in the data and violations of it indicate that for certain strata of the data, a given treatment level is never or rarely experienced. When the ETA is violated estimation methods rely on

extrapolation.

If it is the case that a given treatment level is very rare or non-existent for given strata of the population, an investigator may want to re-consider the original research question of interest. To this end, van der Laan and Petersen (2007) developed causal effect models for realistic intervention rules. These models allow estimation of the effect of realistic interventions, that is only intervening on patients for whom the intervention is reasonably "possible" where "possible" is defined by $g(A \mid W)$ greater than some value, e.g., 0.05. We note that targeted maximum likelihood estimation can be applied to estimate parameters from such models. For applications of such models see Bembom and van der Laan (2007).

The ETA assumption and development of realistic causal models are simply examples of some of the many considerations that arise with observational data as compared to RCT data. However despite the many issues the rich field of causal inference provides promising methods for safety analysis in post-market data. As it is not possible to observe all AE in the pre-market phase, post-market safety analysis is an important and emerging area of research.

12 Discussion

Safety analysis is an important aspect in new drug approvals and has become increasingly evident with the recent cases of drugs withdrawn from the market (e.g., Vioxx). Increasing estimation efficiency is one area that can help overcome the issue that RCT are not powered for safety. Using covariate information is a promising approach to help detect AE that may have remained undetected with the standard crude analysis. Furthermore, time-to-event methods for AE analysis may be more appropriate particularly in studies where the AE often occur for all patients, such as oncology studies. Exploiting the time-dependent nature can further provide more efficient estimates for the effect of treatment on AE occurrence.

In this paper we provided a method for covariate adjustment in RCT for estimating the effect of treatment on the AE failing to occur by a fixed end point. The method does not require any parametric modeling assumptions under MCAR censoring and thus is robust to mis-specification of the hazard fit. The methods advantages were twofold. The first is the potential efficiency gains over the unadjusted estimator. The second is that the targeted maximum likelihood estimator accounts for informative censoring through inverse weighting of the covariate(s) that is added to an initial hazard fit. The standard unadjusted estimator is biased in the informative censoring setting.

The estimator has a relatively straightforward implementation. Given an initial hazard fit either logistic for discrete failure times or Cox proportional hazards for continuous survival times, one updates this fit by iteratively adding a time dependent covariate(s).

The simulation study demonstrated the potential gains in efficiency that can be achieved in addition to the relation of the correlation between the covariate(s) and failure time and efficiency gains. When no predictive covariates were present the relative efficiency was approximately one indicating that one is protected from actually losing precision from applying this method even when the covariates provide little information about failure time. The simulations also demonstrated the reduction in bias in the informative censoring setting.

Considerations for balancing the potential for false positives and the danger of missing possibly significant AE are an important aspect of safety analysis. The strategies from the rich field of multiple testing briefly discussed in this paper can exploit the correlation of the AE outcomes and thus provide the most powerful tests.

While this paper focused on estimation of treatment specific survival at a specific end point an overall average effect of treatment over time may be of interest. The targeted maximum likelihood estimation procedure described in this paper can be extended to estimate this effect to provide a competitor to the ubiquitous log-rank test. Future work includes providing a method for exploiting covariate information using the targeted maximum likelihood estimation procedure to improve power over the log-rank test.

References

Bembom, O. and van der Laan, M. J. (2007). Analyzing sequentially randomized trials based on causal effect models for realistic individualized treatment rules. Technical Report 216, Division of Biostatistics, University of California, Berkeley.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300.

Benjamini, Y., Hochberg, Y., and Kling, Y. (1997). False discovery rate control in multiple hypotheses testing using dependent test statistics. Technical Report 97-1, Tel Aviv University.

Bross, I. D. (1985). Why proof of safety is much more difficult than proof of hazard. *Biometrics*, 31(3):785–793.

Davis, B. R., Furberg, C. D., and Williams, C. B. (1987). Survival analysis of adverse effects data in the beta-blocker heart attack trial. *Clin Pharmacol Ther*, 41:611–615.

Edlavitch, S. A. (1988). Adverse drug event reporting. improving the low us reporting rates. *Arch Intern Med*, 148:14991503.

Fleming, T. R. and Lin, D. Y. (2000). Survival analysis in clinical trials: Past developments and future directions. *Biometrics*, 56(4):971–983.

Friedman, L. M., Furberg, C. D., and DeMets, D. L. (1998). *Fundamentals of Clincal Trials*. Springer-Verlag, New York, 3rd edition.

Gait, J. E., Smith, S., and Brown, S. L. (2000). Evaluation of safety data from controlled clinical trials: the clinical principles explained. *Drug Information Journal*, 34:273–287.

Glasser, S. P., Salas, M., and Delzell, E. (2007). Importance and challenges of studying marketed drugs: What is a phase iv study? common clinical research designs, registries, and self-reporting systems. *J. Clin. Pharmacol.*, 47(9):1074–1086.

Güttner, A., Kübler, J., and Pigeot, I. (2007). Multivariate time-to-event analysis of multiple adverse events of drugs in integrated analyses. *Statistics in medicine*, 26(7):1518–1531.

Hernández, A. V., Eijkemans, M. J., and Steyerberg, E. W. (2006). Randomized controlled trials with time-to-event outcomes: How much does prespecified covariate adjustment increase power? *Annals of Epidemiology*, 16(1):41 – 48.

Hernández, A. V., Steyerberg, E. W., and Habbema, J. D. (2004). Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *J Clin Epidemiol*, 57(5):454–460.

ICH (1996). Harmonised tripartite guideline structure and content of clinical study reports e3. *61 Federal Register*, 37320.

Jacobson, R. M., Adedunni, A., Pankratzc, V. S., and Poland, G. A. (2001). Adverse events and vaccination-the lack of power and predictability of infrequent events in pre-licensure study. *Vaccine*, 19:2428–2433.

Kaplan, K. M., Rusche, S. A., Lakkis, H. D., Bottenfield, G., Guerra, F. A., Guerrero, J., Keyserling, H., Felicione, E., Hesley, T. M., and Boslego, J. W. (2002). Post-licensure comparative study of unusual high-pitched crying and prolonged crying following comvax(tm) and placebo versus pedvaxhib(tm) and recombivax hb(tm) in healthy infants. *Vaccine*, 21(3-4):181 – 187.

Kooperberg, C., Stone, C. J., and Truong, Y. K. (1995). Hazard regression. *Journal of the American Statistical Association*, 90:78–94.

Li, Z. (2001). Covariate adjustment for non-parametric tests for censored survival data. *Statistics in Medicine*, 20(12):1843–1853.

Liu, F. G., Wang, J., Liu, K., and Snavely, D. B. (2006). Confidence intervals for an exposure adjusted incidence rate difference with applications to clinical trials. *Statistics in medicine*, 25(8):1275–1286.

Lu, X. and Tsiatis, A. A. (2008). Improving the efficiency of the log-rank test using auxiliary covariates. *Biometrika*, 95(3):679–694.

Mehrotra, D. V. and Heyse, J. F. (2004). Use of the false discovery rate for evaluating clinical safety data. *Statistical Methods in Medical Research*, 13(3):227–238.

Menjoge, S. S. (2003). On estimation of frequency data with censored observations. *Pharmaceutical Statistics*, 2(3):191–197.

Moore, K. L. and van der Laan, M. J. (2009). Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *Statistics in Medicine*, 28(1):39–64.

Neugebauer, R. and van der Laan, M. J. (2005). Why prefer double robust estimators in causal inference? *Journal of the American Statistical Association*, 129:405–426.

Nishikawa, M., Tango, T., and Ogawa, M. (2006). Non-parametric inference of adverse events under informative censoring. *Statistics in medicine*, 25(23):3981–4003.

O'Neill, R. T. (1988). The assessment of safety. In Peace, K., editor, *Biopharmaceutical Statistics for Drug Development*. Marcel Dekker, New York.

Peace, K. (1987). Design, monitoring and analysis issues relative to adverse events. *Drug Information Journal*, 21(1):21–28.

Pocock, S. J., Assmann, S. E., Enos, L. E., and Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine*, 21:29172930.

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16:285–319.

Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology*, Methodological issues. Bikhäuser.

Robins, J. M. and Rotnitzky, A. (2005). Inverse probability weighted estimation in survival analysis. In Armitage, P. and Colton, T., editors, *Encyclopedia of Biostatistics*. Wiley & Sons, New York, second edition.

Robinson, L. D. and Jewell, N. P. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review*, 59:227240.

Sinisi, S. and van der Laan, M. J. (2004). The deletion/substitution/addition algorithm in loss function based estimation: Applications in genomics. *Statistical Applications in Genetics and Molecular Biology*, 3(1).

Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in medicine*, 27(23):4658–4677.

van der Laan, M. J. (2008). The construction and analysis of adaptive group sequential designs. Technical Report 232, Division of Biostatistics, University of California, Berkeley.

van der Laan, M. J., Dudoit, S., and Pollard, K. S. (2004). ugmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, 3(1).

van der Laan, M. J. and Petersen, M. L. (2007). Causal effect models for realistic individualized treatment and intention to treat rules. *The International Journal of Biostatistics*, 3(1).

van der Laan, M. J. and Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer, New York.

van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):Article 11.

van der Laan, M. J. and Rubin, D. (2007). A note on targeted maximum likelihood and right censored data. Technical Report 226, Division of Biostatistics, University of California, Berkeley.

Wang, Y., Petersen, M. L., Bangsberg, D., and van der Laan, M. J. (2006). Diagnosing bias in the inverse-probability-of-treatment-weighted estimator resulting from violation of experimental treatment assignment. Technical Report 211, Division of Biostatistics, University of California, Berkeley.

Wood, A. M., White, I. R., and Thompson, S. G. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials*, 1(4):368–376.

Zhang, M., Tsiatis, A. A. A., and Davidian, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, 64(3):707 – 715.

Appendix A

Targeted Maximum Likelihood Estimation: A Gentle Introduction

Targeted Maximum Likelihood Estimation: A Gentle Introduction

Susan Gruber and Mark J. van der Laan
Division of Biostatistics, University of California, Berkeley
sgruber@berkeley.edu, laan@berkeley.edu

Abstract

This paper provides a concise introduction to targeted maximum likelihood estimation (TMLE) of causal effect parameters. The interested analyst should gain sufficient understanding of TMLE from this introductory tutorial to be able to apply the method in practice. A program written in R is provided. This program implements a basic version of TMLE that can be used to estimate the effect of a binary point treatment on a continuous or binary outcome.

1 Causal Inference

The counterfactual framework described in Rubin (1974), provides a basis for defining causal effects, such as the difference in mean outcomes between treatment and control groups, relative risk, etc. These causal parameter definitions refer to a full, unobserved, counterfactual dataset containing outcomes for each subject for all possible treatment assignments. In practice the data we measure only contains an outcome value corresponding to the treatment actually assigned. However, the remaining, unobserved outcome(s), can be estimated from observed data to "fill in" the missing, unobserved values, providing that two assumptions, described next, hold. When they do, subsequent parameter estimation from the estimated full dataset is straightforward.

The first assumption, coarsening at random (CAR) , implies that conditional on measured covariates, treatment assignment is independent of the outcome. The second assumption, the experimental treatment assignment (ETA) assumption, requires that the conditional probability of receiving treatment is bounded away from 0 and 1. In other words, observations within strata defined by W have a probability greater than 0 of receiving treatment at all possible levels of the treatment assignment, $\forall a \in \mathcal{A}, P(A = a|W) > 0$. We use the term "theoretical ETA violation" to describe the situation when this assumption does not hold. A "practical ETA violation" occurs when , $\exists a \in \mathcal{A}, P(A = a|W) < \epsilon$, for some small ϵ, typically ranging between (0.1 and 0.01), depending on the number of observations.

In the case of a theoretical ETA violation the causal parameter of interest is not identifiable without additional model assumptions due to a lack of support in the data. When there is a practical ETA violation the parameter of interest is borderline identifiable. Traditional regression techniques are said to "borrow information" to estimate the parameter of interest, but again, this relies on the untestable assumption that the specified model is correct. On the occasions when this modeling assumption is violated, the estimate is biased and the corresponding variance estimates are overly optimistic. It is well-accepted by statisticians that the model is rarely, if ever, correct. Freedman (2005) provides an interesting overview of this topic. A more realistic, non-model-based, causal effect estimate of a borderline-identifiable parameter is likely to have a much larger variance, reflecting the true level of uncertainty in the data.

2 Causal Effect Estimation

We restrict the discussion to estimating the marginal additive effect of a binary point treatment, A, on outcome Y. Given a full (counterfactual) dataset consisting

of n i.i.d. copies of $O_{full} = (W, Y(1), Y(0))$, where $Y_i(1)$ corresponds to the outcome observed when subject i is assigned to the treatment group ($A_i = 1$) and $Y_i(0)$ corresponds to the outcome observed when subject i is assigned to the control group ($A_i = 0$), we can define our parameter of interest as $\psi_0 = E(Y(1) - Y(0))$, the marginal additive treatment effect.

Given observed data, $O_{obs} = (W, A, Y)$, we estimate ψ_0 as:

$$\hat{\psi} = \psi_n = \hat{E}_W(\hat{E}(Y|A = 1, W) - \hat{E}(Y|A = 0, W)).$$

If the outcome or treatment assignment is missing for some observations, the data structure can be expanded to $O_{obs} = (W, A, \Delta, \Delta Y)$, where $\Delta = 1$ when Y is observed, 0 otherwise. In this setting the definition of ψ_0 remains unchanged, but the parameter is estimated as:

$$\hat{E}_W(\hat{E}(Y|A = 1, W, \Delta = 1) - \hat{E}(Y|A = 0, W, \Delta = 1)),$$

where the outer expectation is over all observations.

Common non-parametric or semi-parametric estimators for this problem include the G-computation estimator (Robins, 1986), the inverse-probability-of-treatment (IPTW) estimator (Hernan et al., 2000; Robins, 2000b), the double robust IPTW estimator (Robins and Rotnitzky, 2001; Robins et al., 2000; Robins, 2000a), and targeted maximum likelihood estimation (TMLE) (van der Laan and Rubin, 2006; van der Laan and Gruber, 2009), also doubly-robust. The next section provides an overview of targeted maximum likelihood estimation. The final section describes companion TMLE software for estimating this parameter, written for the R statistical programming environment (R Development Core Team, 2009). Source code is provided in the appendix, along with data analysis examples.

3 Targeted Maximum Likelihood Estimation

Maximum likelihood estimation fits a model to data, minimizing a global measure, such as mean squared error (MSE). When we are interested in one particular parameter of the data distribution and consider the remaining parameters to be nuisance parameters, we would prefer an estimate that has smaller bias and variance for the targeted parameter, at the expense of increased bias and/or variance in the estimation of nuisance parameters. Targeted maximum likelihood estimation targets the MLE estimate of the parameter of interest in a way that reduces bias. This bias reduction is sometimes accompanied by an increase in the variance of the estimate, but the procedure often reduces variance as well in finite samples. Asymptotically, TMLE is maximally efficient when the model and nuisance parameters are correctly specified.

An orthogonal factorization of the likelihood of the data provides the basis for TMLE estimation.

$$\mathcal{L}(O) \quad = \quad P(Y \mid A, W)P(A \mid W)P(W).$$

We define:

$$\begin{aligned} Q(Y, A, W) &\equiv E(Y \mid A, W), \\ g(A, W) &\equiv P(A \mid W), \end{aligned}$$

where $Q(Y, A, W)$ is estimable from the data, $g(A, W)$ is a nuisance parameter that may further factorize into treatment, missingness, and censoring mechanisms, and the empirical distribution of W is the MLE of $P(W)$. For some applications certain factors of g may be known, (e.g., treatment assignment in RCT data), but estimation from the data is common, and can lead to increased efficiency in some cases even when g is known (Moore and van der Laan, 2007). The TMLE estimator is given by:

$$\psi_n^{TMLE} = \frac{1}{n} \sum_{i=1}^{n} Q_n^*(1, W_i) - Q_n^*(0, W_i).$$

Though this parameter is estimated from the Q portion of the likelihood alone, obtaining $Q_n^*(A, W)$, a targeted estimate of the density, involves estimation of nuisance parameter $g(A, W)$ as well.

Super Learner (van der Laan et al., 2007) provides a machine learning approach to data-adaptive estimation of Q_n^0, an initial estimate of Q. The Deletion/Substitution/Addition (DSA) algorithm described in (Sinisi and van der Laan, 2004; Molinaro and van der Laan, 2004) is a less aggressive data-adaptive approach that searches over a large space of polynomial generalized linear models. Alternatively, given a specified parametric model, Q_n^0 can be estimated using standard regression software.This initial estimate is fluctuated in a manner designed to create the largest change in the targeted parameter of the distribution,

$$Q_n^1 = Q_n^0 + \epsilon h(A, W),$$

where $h(A, W)$, a function of the nuisance parameter, depends on the influence curve of the parameter of interest.

The MLE for ϵ is obtained by regressing Y on $h(A, W)$, with offset $Q_n^0(A, W)$. Note that the magnitude of ϵ determines the degree of perturbation of the initial estimate, and is a direct function of the degree of residual confounding. This targeting step maximizes the change in the parameter of interest, but only to the extent that the estimate is confounded along this dimension. It is important to avoid overfitting Q_n^0, as this minimizes the signal in the residuals needed for bias reduction.

3.1 Inference

TMLE estimators are asymptotically normally distributed with mean $\mu = \psi_0$ and variance σ^2/n, where σ^2 is the variance of the influence curve for $\Psi(Q)$. For the parameter of interest specified above, σ^2 is estimated from the data as:

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} \widehat{IC}^2(O_i), \\
\widehat{IC}(Q_n^*, g, \Psi(Q_n^*)) &= h(A,W)(Y - Q_n^*(A,W)) \\
&\quad + Q_n^*(1,W) - Q_n^*(0,W) - \psi_n(Q_n^*), \\
h &= \frac{\Delta}{P(\Delta = 1 \mid A, W)} \left(\frac{I(A=1)}{g(1,W)} - \frac{I(A=0)}{g(0,W)} \right)
\end{aligned}$$

Ninety-five percent confidence intervals, calculated as $\psi_n(Q_n^*) \pm 1.96\hat{\sigma}/\sqrt{n}$, are theoretically well-grounded, and have been shown to provide good coverage in practice across a wide variety of simulated datasets.

A test statistic can be used to test a null hypothesis of the form $H_0 : \psi_0 = 0$:

$$T = \frac{\psi_n}{\sqrt{\hat{\sigma}^2/n}}$$

3.2 Collaborative targeted maximum likelihood estimation

Theoretical findings outlined in van der Laan and Gruber (2009) indicate that it is not always necessary to adjust for the full g_0 in order to obtain unbiased, efficient results. The double robustness property of TMLE estimators guarantees consistent estimation if at least one of Q_0 or g_0 is estimated consistently. Therefore, when $Q_n^0 = Q_0$, adjusting for g_0 is unnecessary. Similarly, when the initial fit for Q contains no information, $(Q_0 - Q_n^0 = Q_0)$, consistent estimation of g_0 is necessary. When Q_n^0 falls somewhere in the middle of these two extremes, adjusting for an essential subset of g_0 allows maximal bias reduction, since the only remaining bias is the residual confounding in $Q_0 - Q_n^0$. CTMLE builds candidate TMLE estimators indexed by $(Q_n^0, g_{n,k}(Q_n^0))$, and selects among using the penalized cross-validated likelihood.

For each stage one estimator, stage two constructs increasingly non-parametric nuisance parameter estimators, $g_{n,1}, \ldots, g_{n,k}$, leading to construction of k updated estimates, $Q_{n,1}^1, \ldots, Q_{n,k}^1$, and a corresponding series of candidate TMLE estimates $(\psi_{n,1}(Q_{n,1}^1), \ldots, \psi_{n,k}(Q_{n,k}^1))$.

3.2.1 Construction of estimators $\{g_{n,1}, ...g_{n,k}\}$

The nature of the candidate estimators for g varies depending on the goodness of fit of the stage one estimate of Q_n^0. When Q_n^0 poorly estimates Q_0, initial estimates of g closely approximate g_0. When Q_n^0 is a good fit for Q_0, the series of candidate estimators of g grows slowly towards estimation of the full g_0. The collaborative nature of the estimation of g is the key difference between standard TMLE and CTMLE.

Though it is a more complex and time-consuming analysis, CTMLE provides two practical advantages over TMLE. First, collaborative, data-adaptive estimation of g leads to reduced variance in the estimate whenever the machine learning procedure determines that adjustment for the full g_0 is unnecessary.

The second advantage occurs in datasets for which the ETA assumption is violated. When there are ETA violations the standard TMLE estimator described above, and the estimated variance, blow up, signaling the lack of identifiability. The CTMLE procedure attempts to remedy the situation by choosing not to adjust for covariates leading to ETA violations. Whether these covariates confound the relationship between treatment and outcome is not knowable from the data. In any case, the CTMLE algorithm will not select a model for g that contains unnecessary covariates, nor will it select a covariate that causes the variance to blow up. This behavior suggests that it is important to understand the reason behind a covariate's exclusion from the model for g. Interpretability plots show the effect on the estimate and the variance of including these covariates in the model. When there is little change in the estimate, we can conclude that the excluded covariate does not bias the estimate. When there is a large change in the estimate and or the variance, we can conclude that there is an ETA violation, but cannot determine from the data the extent of the bias, or even whether the omitted covariate is a true confounder.

4 Discussion

TMLE is a general methodology that can be applied to estimation of many types of causal effect parameters, including but not limited to those involving point treatment effects, survival analysis, longitudinal data analysis, and genomics data. This very generality, and the flexibility allowed for obtaining estimates of the Q and g portions of the likelihood, can perhaps make it difficult for a researcher to understand exactly how to begin analyzing data using TMLE. We endeavor to include just enough information in this paper to allow an interested analyst to begin. To facilitate the process, a set of functions written in R is provided in the appendix. This code defines an implementation of TMLE that can be used to estimate the marginal effect of a binary point treatment on a continuous or binary outcome, even in the presence of missing data. We hope it, too, provides a gentle introduction to the application of targeted maximum likelihood estimation to the estimation of causal efffects.

References

D.A. Freedman. Linear statistical models for causation: A critical review. *Wiley Encyclopedia of Statistics in Behavioral Science*, 2005.

M. A. Hernan, B. Brumback, and J. M. Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 11(5):561–570, 2000.

A. Molinaro and M.J. van der Laan. Deletion/substitution/addition algorithm for partitioning the covariate space in prediction. Technical report, Division of Biostatistics, University of California, Berkeley, 2004.

K.L. Moore and M.J. van der Laan. Covariate adjustment in randomized trials with binary outcomes. Technical report 215, Division of Biostatistics, University of California, Berkeley, April 2007.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

J. M. Robins and A. Rotnitzky. Comment on the Bickel and Kwon article, "Inference for semiparametric models: Some questions and an answer". *Statistica Sinica*, 11(4):920–936, 2001.

J. M. Robins, A. Rotnitzky, and M.J. van der Laan. Comment on "On Profile Likelihood" by S.A. Murphy and A.W. van der Vaart. *Journal of the American Statistical Association – Theory and Methods*, 450:431–435, 2000.

J.M. Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, 2000a.

J.M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:1393–1512, 1986.

J.M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials (Minneapolis, MN, 1997)*, pages 95–133. Springer, New York, 2000b.

D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ Psychol*, 64:688–701, 1974.

S. Sinisi and M.J. van der Laan. The deletion/substitution/addition algorithm in loss function based estimation: Applications in genomics. *Journal of Statistical Methods in Molecular Biology*, 3(1), 2004.

M.J. van der Laan and S. Gruber. Collaborative double robust penalized targeted maximum likelihood estimation. *The International Journal of Biostatistics*, 2009.

M.J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

M.J. van der Laan, E. Polley, and A. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(25), 2007. ISSN 1.

Appendices

The program provided in Appendix A, written for the R statistical environment, can be used to estimate point treatment effects by calling the function *tmle* with the correct arguments. Examples illustrating several options for calling the function are provided in Appendix B.

The simplest approach is demonstrated in Appendix B, example 1. The *tmle* function is called with arguments, Y, A, W, where Y and A are vectors containing the outcome and treatment assignment, respectively. W is a matrix or dataframe where each column corresponds to a potential confounder. The *tmle* function will use the Deletion-Substitution-Addition (DSA) algorithm to estimate Q, g_A, the treatment mechanism, and g_M, the missingness mechanism. This option requires installation of the DSA package, available from:

http://www.stat.berkeley.edu/ laan/Software/

If the DSA package is not installed, Q is estimated with a main terms regression, using glm.

Alternatively, the user can provide working models or numerical values for estimation of any subset of (Q, g_A, g_M), and the program will estimate any that are not user-supplied, see examples 3 and 4.

A complete list of arguments is shown in table 1, below. Return values are described in the description in the last row of the table.

Table 1: Arguments to function *tmle*. Defaults for optional arguments are listed in parentheses. (*) indicates required argument.

Argument	Description
Y*	Outcome variable, continuous or binary. Missing values allowed.
A*	Binary treatment indicator, 1 - treatment, 0 - control. Missing values allowed.
W*	Baseline covariate,. numerical vector, matrix, or dataframe.
Delta (1 for all obs)	Indicator of missingness for (Y, A), 1 - observed, 0 - missing
id (1 to n)	identify repeated measures
Q (DSA estimation)	$E(Y\|A, W)$, specified in one of three ways: 1. NULL - defaults to DSA estimation of $E(Y\|A = a, W, \Delta)$ 2. matrix of values containing three columns: $(E(Y\|A = a, W, \Delta), E(Y\|A = 1, W, \Delta), E(Y\|A = 0, W, \Delta)$ 3. formula for estimation of $E(Y\|A, W, \Delta)$, to use with glm
g_A (DSA estimation)	$P(A = 1\|W)$, treatment mechanism specified in one of three ways: 1. NULL - defaults to DSA estimation of $P(A = 1\|W)$ 2. vector of values $P(A = 1\|W)$ 3. formula for estimation of $P(A = 1, W)$, to use with glm
g_M (DSA estimation)	$P(\Delta = 1\|W)$, missingness mechanism for (A, Y) specified in one of three ways: 1. NULL - defaults to DSA estimation of $P(\Delta = 1\|W)$ 2. vector of values $P(\Delta = 1\|W)$ 3. formula for estimation of $P(\Delta = 1, W)$, to use with glm
wts (1 for all obs)	weights for observations
DSAargs	a list containing settings for DSA estimation. Defaults: DSAargs$formula = $Y \sim A$, DSAargs$maxsumofpow = 2, DSAargs$maxorderint = 2, DSAargs$vfold = 5, DSAargs$family = gaussian DSAargs$maxsize=min(2*ncol(W),15) (model size capped at 15), DSAargs$nsplits=1, DSAargs$Dmove=TRUE, DSAargs$Smove=TRUE
DETAILED (FALSE)	flag specifying basic or detailed return value. TRUE: psi - treatment effect estimate, var - estimated variance of parameter estimate, epsilon - coefficient used in targeting step, coefficients and predicted values for $Q_n^0(A, W.\Delta), g_A(1, W, \Delta), g_M(1, W, \Delta)$ FALSE: psi- treatment effect estimate, var - estimated variance of parameter estimate

Appendix A: R implementation of TMLE

```
#-------------------------------------------------------------------------------
# Targeted Maximum Likelihood Estimation
# for binary point treatment, non-parametric estimation
# paramter of interest  = E_W[E(Y|A=1,W) - E(Y|A=0,W)]
# taking into account treatment (g_A) and missingness (g_M) mechanisms
# models or estimates for Q, g_A, g_M can be user-supplied or estimated internally using DSA
# as implemented, arguments to DSA are the same for all estimation procedures
# these can be user-supplied or set to default values
# maxorderint = 2, maxsumofpow=2, maxsize = 15
# Dmove=TRUE, Smove=TRUE, formula = Y~A, A forced into model.
#
# August 16, 2009
# Susan Gruber, sgruber@berkeley.edu
#
# for information see
# M.J. van der Laan and D. Rubin. Targeted maximum likelihood learning.
# The International Journal of Biostatistics, 2(1), 2006.
# http://www.bepress.com/ijb/vol2/iss1/11/


#-------------verify_args------------------
verify_args <- function(Y,A,W,Delta){
ok1 <- length(Y) == length(A) & length(A) == nrow(W)
ok2 <- all(A[!is.na(A)] %in% 0:1)
if (!ok1) {warning("Y, A, W must contain the same number of observations")}
if (!ok2) {warning("A must be binary (0,1)")}
return(ok1&ok2)
}

#-----------set_DSAargs----------------
set_DSAargs <- function(DSAargs, wts){
   if(is.null(DSAargs$maxsumofpow)){DSAargs$maxsumofpow <- 2}
   if(is.null(DSAargs$maxorderint)){DSAargs$maxorderint <- 2}
   if(is.null(DSAargs$maxsize)) {DSAargs$maxsize <- 15}
   if(is.null(DSAargs$Dmove)) {DSAargs$Dmove <- TRUE}
   if(is.null(DSAargs$Smove)) {DSAargs$Smove <- TRUE}
   if(is.null(DSAargs$vfold)) {DSAargs$vfold <- 5}
   if(is.null(DSAargs$formula)){DSAargs$formula <- Y~A}
   if(is.null(DSAargs$family)){DSAargs$family <- "gaussian"}
   if(is.null(DSAargs$silent)) {DSAargs$silent <- TRUE}
   if(is.null(DSAargs$wts)) {DSAargs$wts <- matrix(data=rep(wts, DSAargs$vfold+1),
byrow=TRUE, nrow=DSAargs$vfold+1)}
   if(is.null(DSAargs$nsplits)) {DSAargs$nsplits <- 1}
   if(is.null(DSAargs$silent)) {DSAargs$silent <- -1}
   return(DSAargs)
}

#-----------function logit---------
# convert probability to logit
# truncate probability passed in
#------------------------------
logit <- function(x){
        x[x>1] <-1
        x[x<0] <-0
        return(-log(1/x - 1))
}
```

```
#-----------estimate_Q----------------
# figure out if Q is one of three things:
# 1.  a matrix of values, QAW, Q1W, Q0W
# 2.  a model to use glm on
# 3.  null - estimate with DSA if available, otherwise main terms with glm
# returns matrix of linear predictors for Q(A,W), Q(1,W), Q(0,W)
#---------------------------------------
estimate_Q <- function (Q, DSAargs, Y,A,W, Delta, family, wts, id) {
   if(is.matrix(Q)){
     if (family == "binomial") {Q <- logit(Q)}
     coef <- NA
   } else {
     if (is.null(Q)){
   if(require(DSA)){
   DSAargs <- set_DSAargs(DSAargs, wts)
   m <- DSA(formula=DSAargs$formula, data=data.frame(Y,A,W)[Delta==1,],
       weights=DSAargs$wts[,Delta==1], id=id[Delta==1],
   maxsumofpow=DSAargs$maxsumofpow, maxorderint=DSAargs$maxorderint,
   maxsize=DSAargs$maxsize, Dmoves=DSAargs$Dmove, Smove=DSAargs$Smove,
   family=family, candidate.rank=DSAargs$candidate.rank,
   rank.cutoffs = DSAargs$rank.cutoffs, usersplits=DSAargs$usersplits,
   userseed=DSAargs$userseed, vfold=DSAargs$vfold, nsplits=DSAargs$nsplits,
   silent=DSAargs$silent )
   } else {
   warning("DSA not found, running main terms regression for Q using glm")
   form <- paste("Y~A", paste(colnames(W), collapse = "+"), sep="+")
   m <- glm(form, family=family, data=data.frame(Y,A,W, wts, Delta), weights=wts,
na.action=na.exclude, subset=Delta==1)
   }
     } else {
      form <- try(as.formula(Q))
      if(class(form)== "formula") {
      m <- glm(form,  family=family, data=data.frame(Y,A,W, wts, Delta), weights=wts,
na.action=na.exclude, subset=Delta==1)
      } else {
      warning("Invalid formula supplied, running main terms regression for Q using glm")
      form <- paste("Y~A", paste(colnames(W), collapse = "+"), sep="+")
   m <- glm(form, family=family, data=data.frame(Y,A,W, wts, Delta), weights=wts,
 na.action=na.exclude, subset=Delta==1)
   }
     }
     QAW <- predict(m, newdata=data.frame(Y,A,W))
     Q1W <- predict(m, newdata=data.frame(Y,A=1,W))
     Q0W <- predict(m, newdata=data.frame(Y,A=0,W))
     Q <- cbind(QAW, Q1W, Q0W)
     coef <- coef(m)
    }
    return(list(Q=Q, coef=coef))
}

#-----------estimate_g----------------
# Estimate any factor of g
#---------------------------------------
estimate_g <- function (g, DSAargs,A,W, Delta, wts, id) {
  if (!is.numeric(g)){
if (all(A==A[1])) {
g1W <- 1
coef<- NA
} else {
```

```
    if (is.null(g)){
   if(require(DSA)){
   DSAargs <- set_DSAargs(DSAargs, wts)
   m <- DSA(formula=DSAargs$formula, data=data.frame(A,W)[Delta==1,],
       weights=DSAargs$wts[,Delta==1], id=id[Delta==1],
   maxsumofpow=DSAargs$maxsumofpow, maxorderint=DSAargs$maxorderint,
   maxsize=DSAargs$maxsize, Dmoves=DSAargs$Dmove, Smove=DSAargs$Smove,
   family="binomial", candidate.rank=DSAargs$candidate.rank,
   rank.cutoffs = DSAargs$rank.cutoffs, usersplits=DSAargs$usersplits,
   userseed=DSAargs$userseed, vfold=DSAargs$vfold, nsplits=DSAargs$nsplits,
   silent=DSAargs$silent )
   } else {
   warning("DSA not found, running main terms regression for g using glm")
form <- paste("A~1", paste(colnames(W), collapse = "+"), sep="+")
m <- glm(form, family="binomial", data=data.frame(A,W, wts, Delta), weights=wts,
na.action=na.exclude, subset=Delta==1)
   }
     } else {
      form <- try(as.formula(g))
      if(class(form)== "formula") {
      m <- try(glm(form,  family="binomial", data=data.frame(A,W, wts, Delta), weights=wts,
 na.action=na.exclude, subset=Delta==1))
      if (class(m)[1]=="try-error"){
      warning("Invalid formula supplied, running main terms regression for g using glm")
      form <- paste("A~1", paste(colnames(W), collapse = "+"), sep="+")
      m <- glm(form, family="binomial", data=data.frame(A,W, wts, Delta),
weights=wts,na.action=na.exclude, subset=Delta==1)
      }
      } else {
      form <- paste("A~1", paste(colnames(W), collapse = "+"), sep="+")
      m <- glm(form, family="binomial", data=data.frame(A,W, wts, Delta), weights=wts,
na.action=na.exclude, subset=Delta==1)
      }
     }
     g1W <- predict(m, newdata=data.frame(A,W,wts), type="response")
     coef <- m$coef
   }
 } else {
  g1W <- g
  coef <- NA
 }
 return(list(g1W=g1W, coef=coef))
}

#------------------------------tmle----------------------------------------
# estimate marginal treatment effect for binary point treatment
# accounting for missing outcomes.
# arguments:
# Y - outcome
# A - binary treatment indicator, 1-treatment, 0 - control
# W - vector, matrix or dataframe containing baseline covariates
# Delta - indicator of missing outcome or treatment assignment.  1 - observed, 0 - missing
# id - id identifying repeated measures
# Q - E(Y|A,W), specified in one of three ways:
# 1. NULL - defaults to DSA estimation of E(Y|A=a, W), with A forced into the model
# 2. matrix of values containing three columns.  1: E(Y|A=a,W), 2: E(Y|A=1,W), 3: E(Y|A=0,W)
# 3. formula for estimation of E(Y|A, W), suitable for call to glm
# g_A - binary treatment mechanism, specified in one of three ways:
# 1. NULL - defaults to DSA estimation of P(A=1|W)
# 2. vector of values P(A=1|W)
```

```
# 3. formula for estimation of P(A=1,W), suitable for call to glm
# g_M - missingness mechanism, specified in one of three ways:
# 1. NULL - defaults to DSA estimation of P(Delta=1|W)
# 2. vector of values P(Delta=1|W)
# 3. formula for estimation of P(Delta=1,W), suitable for call to glm
# wts - optional weights on observations
# DSAargs - optional settings for DSA estimation
#  defaults:  maxsumofpow = 2, maxorderint = 2, maxsize=min(2*ncol(W),15) (model size capped at 15),
#  vfold = 5, nsplits=1, Dmove=TRUE, Smove=TRUE
# family - family specification for regression models, defaults to gaussian
# DETAILED - flag indicating basic or detailed return value.
# TRUE  - psi, treatment effect estimate,
#    var - estimated variance of parameter estimate,
#    epsilon - coefficient used in targeting step
#    coefficients and predicted values for Q_n^0(A,W), g_A(1,W), g_M(1,A,W)
# FALSE - psi, treatment effect estimate,
#    var - estimated variance of parameter estimate
#-------------------------------------------------------------------------------

tmle <- function(Y,A,W,Delta=rep(1,length(Y)), id=1:length(Y), Q=NULL, g_A=NULL, g_M=NULL,
wts=rep(1, length(Y)), DSAargs=NULL, family="gaussian", DETAILED=FALSE) {
psi.tmle <- varIC <- NA
W <- as.matrix(W)
if(verify_args(Y,A,W,Delta)){
Q <- estimate_Q(Q, DSAargs, Y,A,W, Delta, family, wts, id)
DSAargs$formula <- A~1
    g_A <- estimate_g(g_A, DSAargs, A, W, Delta, wts, id)
    g_M <- estimate_g(g_M, DSAargs, A=Delta, W, Delta=rep(1,nrow(W)), wts, id)
   g1W <- g_A$g1W
   h <- h1W <- 1/g1W * Delta/g_M$g1W
   h0W <- -1/(1-g1W) *  Delta/g_M$g1W
   h[A==0] <- h0W[A==0]
   epsilon <- coef(glm(Y~-1 + offset(Q$Q[,1]) + h, family=family, weights=wts, subset=Delta==1))

   QAW <- Q$Q[,1] + epsilon*h
   Q1W <- Q$Q[,2] + epsilon*h1W
   Q0W <- Q$Q[,3] + epsilon*h0W

   if (identical(family, binomial) | identical(family,"binomial")) {
   QAW <- 1/(1+exp(-QAW))
   Q1W <- 1/(1+exp(-Q1W))
   Q0W <- 1/(1+exp(-Q0W))
   }
   psi.tmle <- mean(Q1W) - mean(Q0W)
   Y[is.na(Y)] <- QAW[is.na(Y)] # keeps arithmetic from failing
   IC <- (Y-QAW)*h*Delta + Q1W - Q0W - psi.tmle
     IC <- as.vector(by(IC, id, sum))
     IC[is.nan(IC)|is.infinite(IC)] <- Inf
     varIC <- var(IC)
   }
   if (DETAILED) {
Qcounter <- cbind(Q1W, Q0W)
colnames(Qcounter) <- c("Q1W", "Q0W")
   returnVal <- list(psi=psi.tmle, var = varIC/length(unique(id)),epsilon=epsilon, Q=Q,
g_A=g_A, g_M=g_M, Qcounter=Qcounter)
   } else {
   returnVal <- list(psi=psi.tmle, var = varIC/length(unique(id)))
   }
   return(returnVal)
}
```

Appendix B: Sample calls to tmle function

```
#-------------------------------------------------------------------------------
# tmle examples
# use with function tmle in file tmle.R
# Susan Gruber
# sgruber@berkeley.edu
# August 16, 2009

# Important:  Generate data before running the examples!
# psi_0 = 1

#------------generate data --------------

  set.seed(10)
  n <- 500
  W <- matrix(rnorm(n*3), ncol=3)
  A <- rbinom(n,1, 1/(1+exp(-(.1*W[,1] - .1*W[,2] + .5*W[,3]))))
  Y <- A + 2*W[,1] + W[,3] + W[,2]^2 + rnorm(n)

  colnames(W) <- paste("W",1:3, sep="")

#-----------------------------------------------------
# Example 1, default function invocation
# invokes DSA to estimate Q, g_A, g_M,
# because Delta argument is not supplied, assumes (Y,A) observed for all obs

  result1 <- tmle(Y,A,W)

#-----------------------------------------------------
# Example 2: Binary outcome, DSA estimates Q
# known g_A = 0.5 is user-supplied,
#
  A.ex2 <- rbinom(n,1,.5)
  Y.ex2 <- A.ex2 + 2*W[,1] + W[,3] + W[,2]^2 + rnorm(n)
  result2 <- tmle(Y=Y.ex2,A=A.ex2,W, g_A =rep(.5, length(Y)))

#-----------------------------------------------------
# Example 3: Supplying an indicator for observations missing the outcome
# set Delta to 1 for obs where Y is observed, 0 when Y is missing
# In this example, Delta is set to indicate 20% missing values, MCAR
# DSA to estimate Q, g_A, g_M,
# set DETAILED=TRUE to see model selected by DSA and predicted values
# for Q_n^0, g_A, g_M for each observation, and epsilon.

  Delta <- rbinom(n,1,.8)
  result3 <- tmle(Y,A,W, Delta=Delta, DETAILED=TRUE)

#-----------------------------------------------------
# Example 4: User-supplied (misspecified) model for Q, DSA estimates for g_A, g_M
# approx. 20% missing, MAR

  Delta <- rbinom(n, 1, 1/(1+exp(-(1.7-1*W[,1]))))
  result4 <- tmle(Y,A,W, Delta=Delta, Q=Y~A+W1+W2+W3, DETAILED=TRUE)

#-----------------------------------------------------
# Example 5: User-supplied models for g_A and missingness mechanism g_M,
#  DSA estimates Q.
#  100 unique IDs supplied
# Usage note: use "A" for dependent variable name in the formula for g_M
```

```
  Delta <- rbinom(n, 1, 1/(1+exp(-(1.6-1*W[,1]))))
  result5 <- tmle(Y,A,W, Delta=Delta, g_A=A~W1+W2+W3, g_M=A~W1, id=rep(1:100, length=n), DETAILED=TRUE)

#-------------------------------------------------------
results_summary <- cbind(c(result1$psi, result2$psi, result3$psi, result4$psi, result5$psi),
          c(result1$var, result2$var, result3$var, result4$var, result5$var))

colnames(results_summary) <- c("estimate", "variance")
print(results_summary,digits=3)
```

Appendix B

Targeted Maximum Likelihood Learning: Examples and Generalizations

The following work is currently unpublished elsewhere.

Targeted Maximum Likelihood Learning: Examples and Generalizations

Mark J. van der Laan
Division of Biostatistics, University of California, Berkeley
laan@stat.berkeley.edu

Abstract

This paper should be read as a follow up on our general targeted maximum likelihood learning article (van der Laan, Rubin, 2006). In the current paper I present some additional applications and thereby illustrations of the general targeted maximum likelihood methodology (van der Laan, Rubin, 2006). These examples illustrate how targeted maximum likelihood learning can be used effectively to efficiently estimate causal effects of a treatment on an outcome of interest based on observational as well as clinical trial data, and to efficiently estimate variable importance parameters measuring the importance of a variable (e.g., biomarker) in predicting an outcome, while adjusting for a set of other variables, based on censored and uncensored data. Targeted maximum likelihood estimation of variable importance parameters has important applications in genomics and biomarker discovery, among many others.

In addition, we present the analogue of the iterative targeted maximum likelihood estimator presented in (van der Laan, Rubin, 2006) to a Baysian setting resulting in a targeted posterior distribution on the parameter of interest, given a prior distribution on this parameter of interest. We also present some obvious generalizations of the targeted maximum likelihood learning methodology by (e.g.,) replacing the log likelihood loss function by any other loss function.

Key words: Causal effect, efficient influence curve, estimating function, locally efficient estimation, loss function, maximum likelihood estimation, posterior distribution, targeted maximum likelihood estimation, variable importance.

1 Targeted MLE for realistic Marginal structural models.

The observed data structure on each experimental unit is $O = (W, A, Y)$, where W is a collection of baseline covariates, A is a treatment variable, and Y is an outcome of interest. We observe n i.i.d. copies $O_1, \ldots, O_n$, and the goal is to estimate the causal effect of treatment on the outcome within subgroups defined by the strata of a baseline covariate V included in W. This has important applications in causal effect estimation of a drug (e.g. dose) in clinical trials as well for observational (e.g post market) studies.

The full data structure and parameter of interest: Let $Y(a)$ represent a treatment specific outcome one would observe if the randomly sampled subject would be assigned a treatment coded as $a \in \mathcal{A}$, and let $X = (W, (Y(a) : a \in \mathcal{A})) \sim P_{X0}$ represent the full data structure of interest on the randomly sampled subject consisting of the treatment specific outcomes, and baseline covariates W. Let $\mathcal{A}_1$ denote an index set of a set of dynamic point treatment rules

$$\mathcal{D} = \{W \to d(a)(W) \in \mathcal{A} : a \in \mathcal{A}_1\},$$

where each rule in this set $\mathcal{D}$ of rules, represents a rule for assigning treatment in response to the subject's/experimental unit's baseline covariates W. A special case is that $\mathcal{A}_1 = \mathcal{A}$ and $d(a)$ denotes a rule which aims to assign a but if a is such that the conditional probability $g_0(a \mid W)$ of treatment being equal to a, given the baseline covariates W, is too close to zero, then it assigns a treatment in the set $\mathcal{A}$ of possibly treatment options closest to a, where the latter "closest" needs to be defined appropriately. We refer to such rules avoiding treatment assignment which are not supported by the treatment mechanism g_0 as realistic treatment rules.

We consider a model in which the full data distribution P_{X0} is unspecified. Let $X_1, \ldots, X_n$ be n i.i.d. draws of X. A scientific parameter of interest is a realistic causal treatment curve defined as the mean $\psi_0(a) = E_0 Y(d(a))$ of the treatment specific outcome $Y(d(a))$, where $d(a)$ is a dynamic point treatment rule $W \to d(a)(W)$, and $Y(d(a))$ represents the outcome one would observe if the subject follows this rule. In addition, we are also concerned with the V-adjusted causal response curve for a $V \subset W$ defined as

$$\psi_0(a, v) = E_0(Y(d(a)) \mid V = v),$$

where V represents a baseline characteristic which might potentially strongly affect the causal response curve.

Here $d(a)$ is a dynamic point treatment rule $W \to d(a)(W)$ mapping the baseline covariates in the set $\mathcal{A}$ of treatment options satisfying for some user supplied $\delta > 0$ the following condition:

$$P(A = d(a)(W) \mid W) > \delta \text{ almost everywhere, for all } a \in \mathcal{A}_1. \quad (1)$$

A counterfactual $Y(d(a))$ indexed by a dynamic treatment rule $d(a)$ is a well defined function of the complete set of counterfactuals $(Y(a) : a \in \mathcal{A})$ and baseline covariates W, and the rule $d(a)$: $Y(d(a)) = Y(d(a)(W))$.

Missing data structure representation of observed data on experimental unit: It is assumed that $O = (W, A, Y = Y(A))$ with probability 1.
Randomization assumption: We also assume that A is randomized conditional on W:

$$g_0(a \mid X) = P(A = a \mid X) = P(A = a \mid W).$$

The assumption (1) guarantees that the distribution of the counterfactual $Y(d(a))$ is identifiable from the observed data structure $O = (W, A, Y = Y(A))$.

Working model: We consider a working model $m(a, v \mid \beta)$ for the treatment specific mean $\psi_0(a, v)$, and define the target parameter as

$$\beta_0 = \arg\min_{\beta} E_{0V} \sum_{a \in \mathcal{A}_1} (m(a, V \mid \beta) - \psi_0(a, V))^2 h(a, V),$$

where h is a user supplied weight function. For simplicity, we assume here that $\mathcal{A}_1$ is discrete, but if $\mathcal{A}_1$ is a continuous set, then one can replace it by a discrete approximation in the above definition.

The summary measure $\tilde{\psi}_0(a, v) = m(a, v \mid \beta_0)$ of ψ_0 implied by the working model $\{m(\cdot \mid \beta) : \beta\}$ provides now a model based approximation of the true causal response curve ψ_0. Note that β_0 is a parameter of ψ_0 and the marginal distribution P_{0V} of V. Although, we will consider the model for the full data distribution P_{X0} to be nonparametric and the working model as an approximation of the true causal response curve, our proposed estimators are valid if one actually assumes the working model $m(a, V \mid \beta_0)$ to be correctly specified.Thus our goal is to construct a targeted MLE of β_0.

Important identity: Under a mild regularity condition, it follows that

$\beta_0 = \beta(Q_{01}, Q_{02})$ solves

$$\begin{aligned} 0 &= E_{0V} \sum_{a \in \mathcal{A}_1} h(a,V) \frac{d}{d\beta_0} m(a,V \mid \beta_0)(E(Y(d(a)) \mid V) - m(a,V \mid \beta_0)) \\ &= E_{Q_{01}} \sum_{a \in \mathcal{A}_1} h(a,V) \frac{d}{d\beta_0} m(a,V \mid \beta_0)(E(Y(d(a)) \mid W) - m(a,V \mid \beta_0)) \\ &= E_{Q_{01}} \sum_{a \in \mathcal{A}_1} h(a,V) \frac{d}{d\beta_0} m(a,V \mid \beta_0)(Q_{02}(d(a),W) - m(a,V \mid \beta_0)), \end{aligned}$$

where we defined $Q_{02}(d(a), W) = E(Y \mid A = d(a)(W), W)$. This identity will be assumed to hold.

Optimal treatment: We are also concerned with statistical inference for the optimal treatment for subgroup v

$$a^*(\beta_0)(v) = \arg\max_{a \in \mathcal{A}_1} m(a, v \mid \beta_0),$$

and, in case V is chosen to be the empty set, then this reduces to the marginal optimal treatment

$$a^*(\beta_0) = \arg\max_{a \in \mathcal{A}_1} m(a \mid \beta_0).$$

A particular working model of interest for determining an optimal treatment among a continuous set $\mathcal{A}_1$ is given by a quadratic model

$$m(a, v \mid \beta_0) = \beta_0(0)(v) + \beta_0(1)(v)a + \beta_0(2)(v)a^2,$$

where, for example, $\beta_0(j)(v) = \beta_0(j)(0) + \beta_0(j)(1)v$, $j = 0, 1, 2$. Such a quadratic model allows for applications in which the optimal dose is neither the maximum value nor the minimum, but something in between. For this choice of working model we have that the optimal dose for subgroup $V = v$ is given by:

$$a^*(\beta_0)(v) = \frac{-\beta_0(1)(v)}{2\beta_0(2)(v)}.$$

In particular, the optimal marginal dose is given by

$$a^*(\beta_0) = \frac{-\beta_0(1)}{2\beta_0(2)}.$$

Likelihood and Identifiability: Firstly, we note that the likelihood of the observed data set $(O_1, \ldots, O_n)$ factorizes as:

$$P_{Q_0,g}(O_1, \ldots, O_n) = \prod_{i=1}^n Q_{10}(W_i) Q_{20}(Y_i \mid A_i, W_i) \prod_{i=1}^n g(A_i \mid W_i),$$

where the conditional density of Y_i, given $A_i = a$, W_i, $Q_{20}(\cdot \mid a, W_i)$, equals the conditional density of $Y_i(a)$, given W_i, and Q_{10} denotes the marginal density of W. In particular, it follows that the marginal causal dose response curve $\psi_0(a)$ is identified by the Q_0-factor of the likelihood by the following relation:

$$\psi_0(a) = E_0 E_0(Y \mid A = d(a)(W), W).$$

In general, under this same condition,

$$\psi_0(a, v) = E_0\{E_0(Y \mid A = d(a)(W), W) \mid V = v\}.$$

Maximum Likelihood Estimation: Consider a model $\{Q_{2\theta} : \theta\}$ for the distribution of $Y(a)$, given W, or equivalently, the distribution Q_{02} of Y, given A, W, and the corresponding maximum likelihood estimator θ_n:

$$\theta_n = \arg\max_\theta \sum_{i=1}^n \log Q_{2\theta}(Y_i \mid A_i, W_i).$$

We will leave the marginal distribution of W unspecified, so that this is estimated with the empirical probability distribution Q_{1n} of $W_1, \ldots, W_n$. The model $\{Q_{2\theta} : \theta\}$ defines a working model $\mathcal{Q}^w$ for the unknown components $Q_0 = (Q_{10}, Q_{20})$ of the likelihood of the observed data. Given an estimator θ_n, we will use the short-hand notation $Q_{\theta_n} = (Q_{1n}, Q_{2\theta_n})$ for the estimate of both the marginal distribution of W as well as the conditional distribution of Y, given A, W. We also assume that we are given an estimate g_n of the treatment mechanism $g_0(A \mid W)$ in the case that the latter is not known by design.

We wish to compute the targeted MLE for the nonparametric model targeting β_0, based on an initial maximum likelihood estimator Q_{θ_n} based on this working model $\mathcal{Q}^w$. For this purpose, we first need to know the efficient influence curve of β_0 in our nonparametric model for the observed data O.

Efficient influence curve: The efficient influence curve for β_0 at P_{Q_0,g_0} is, up till a normalizing matrix, given by

$$\begin{aligned} D^*(Q_0, g_0) &= \sum_{a \in \mathcal{A}_1} I(A = d(a)(W)) \frac{h(a, V) \frac{d}{d\beta_0} m(a, V \mid \beta_0)}{g_0(A \mid X)} (Y - Q_{02}(A, W)) \\ &\quad + \sum_{a \in \mathcal{A}_1} h(a, V) \frac{d}{d\beta_0} m(a, V \mid \beta_0)(Q_{02}(d(a), W) - m(a, V \mid \beta_0)) \\ &\equiv D_1^*(Q_0, g_0)(W, A, Y) + D_2^*(Q_0)(W), \end{aligned}$$

where we defined $Q_{02}(d(a), W) = E_{Q_0}(Y \mid A = d(a)(W), W)$ and $Q_{02}(a, W) = E(Y \mid A = a, W)$, and we note that $\beta_0 = \beta(Q_0)$ is a parameter of $Q_0 =$

(Q_{01}, Q_{02}). The IPTW component of $D^*(Q_0, g_0)$ is $D_{IPTW}(g_0, \beta_0) = \sum_{a \in \mathcal{A}_1} I(A = d(a)(W)) \frac{h(a,V) \frac{d}{d\beta_0} m(a,V|\beta_0)}{g_0(A|X)} (Y - m(a, V \mid \beta_0))$ and we have the usual DR-IPTW representation $D^* = D_{IPTW} - E(D_{IPTW} \mid A, W) + E(D_{IPTW} \mid W)$ of D^*.

Let

$$\begin{aligned} c(P_{Q_0,g_0}, g_0, \beta_0) &= P_{Q_0,g_0} \sum_{a \in \mathcal{A}_1} I(A = d(a)(W)) \frac{h(a,V)}{g_0(A \mid X)} \frac{d}{d\beta_0} m(a, V \mid \beta_0) \frac{d}{d\beta_0} m(a, V \mid \beta_0)^\top \\ &= E_{Q_0} \sum_{a \in \mathcal{A}_1} h(a,V) \frac{d}{d\beta_0} m(a, V \mid \beta_0) \frac{d}{d\beta_0} m(a, V \mid \beta_0)^\top. \end{aligned}$$

The efficient influence curve for β_0 is given by $c(P_{Q_0,g_0}, g_0, \beta_0)^{-1} D^*(Q_0, g_0)$. The efficient influence curve for a (e.g. lower dimensional) function of β_0 can be derived (as a linear mapping applied to the vector efficient influence curve D^*) based on the δ-method. The targeted MLE could be equally well developed for this function by the efficient influence curve of the lower dimensional function instead, possibly up till a normalizing matrix. Below, we present the targeted MLE for the whole β_0.

Epsilon-fluctuation for Targeted MLE: Let $\{Q_{2\theta}(\epsilon) : \epsilon\}$ be a path through $Q_{2\theta}$ at $\epsilon = 0$ and satisfy the score condition $\frac{d}{d\epsilon} \log Q_{2\theta}(\epsilon)\big|_{\epsilon=0} = D_1^*(Q_{2\theta}, g_0)$. (For the targeted MLE for functions of β_0 we would also decompose its efficient influence curve in a D_1^* component representing its projection on functions of O with conditional mean zero, given A, W, and D_2^* component representing its projections on the functions of W with mean zero). For example, if $Q_{2\theta}$ is a regression model of Y on A, W with normal errors with constant variance, then we can simply add the extension $\epsilon C^*(A, W)$, where

$$C^*(A, W) \equiv \sum_{a \in \mathcal{A}_1} I(A = d(a)(W)) \frac{h(a,V) \frac{d}{d\beta_0} m(a, V \mid \beta_0)}{g_0(A \mid X)}.$$

In other words, $E_{Q_{2\theta}(\epsilon)}(Y \mid A, W) = E_{Q_{2\theta}}(Y \mid A, W) + \epsilon C^*(A, W)$.

Similarly, if $Q_{2\theta}$ is a logistic regression of a binary Y on A, W, then we simply add $\epsilon \sum_{a \in \mathcal{A}_1} \frac{h(a,V) \frac{d}{d\beta_0} m(a,V|\beta_0)}{g_0(A|X)}$ to the logit of $Q_{2\theta}(1 \mid A, W)$. In other words,

$$\text{logit} E_{Q_{2\theta}(\epsilon)}(Y \mid A, W) = \text{logit} E_{Q_{2\theta}}(Y \mid A, W) + \epsilon C(A, W).$$

In both cases, these ϵ extensions have a score at $\epsilon = 0$ equal to $D_1^*(Q_{2\theta}, g_0)$.

Making the epsilon-covariate extension independent of β_0: The targeted MLE can be obtained in one maximum likelihood step determining

the maximum likelihood estimator of ϵ in the case that the epsilon-covariate $C(A,W)$ does not depend on β_0. In the case that $m(a,V \mid \beta)$ is a linear regression model, say $m(a,V \mid \beta) = \beta(a,V)$, then $h(a,V)\frac{d}{d\beta_0}m(a,V \mid \beta_0) = h(a,V)(a,V)^\top$ so that indeed the ϵ-covariate is independent of β_0 for each choice of h.

In the case that $m(a,V \mid \beta)$ is a logistic linear regression model, say, $m(a,V \mid \beta_0) = 1/(1+\exp(-\beta_0(a,V))$, then we recommend to select $h(a,V) = h_1(a,V)/(m(a,V \mid \beta_0)(1-m(a,V \mid \beta_0))$ for some h_1 so that $h(a,V)\frac{d}{d\beta_0}m(a,V \mid \beta_0)$ reduces to $h_1(a,V)(a,V)^\top$ and is thus independent of β_0. Similarly, if $m(a,V \mid \beta)$ is a log linear regression model (modelling a causal relative risk), say $m(a,V \mid \beta) = \exp(\beta(a,V))$, then we could select $h(a,V) = h_1(a,V)/m(a,V \mid \beta_0)$ for some h_1 so that $h(a,V)\frac{d}{d\beta_0}m(a,V \mid \beta_0)$ reduces to $h_1(a,V)(a,V)^\top$ so that the ϵ-covariate is thus independent of β_0 again.

The one-step targeted MLE: Given an estimate g_n of the treatment mechanism g_0, let ϵ_n be the solution of

$$0 = \sum_i D_1^*(Q_{2\theta_n}(\epsilon_n), g_n)(O_i).$$

In the above two linear and logistic regression ϵ-extensions, and under the assumption that the ϵ-extension covariate $C(A,W) = \sum_{a\in\mathcal{A}_1} \frac{h(a,V)\frac{d}{d\beta_0}m(a,V|\beta_0)}{g_n(A|X)}$ does not depend on β_0, it follows that

$$\epsilon_n = \arg\max_\epsilon \sum_{i=1}^n \log Q_{2\theta_n}(\epsilon)(O_i)$$

is the maximum likelihood estimator over ϵ.

We call $\beta_n = \beta(Q_{1n}, Q_{2\theta_n}(\epsilon_n))$ corresponding with the updated $Q_{\theta_n}(\epsilon_n)$ the targeted MLE of β_0. Recall the above mentioned identity

$$0 = E_{Q_{01}} \sum_{a\in\mathcal{A}_1} h(a,V)\frac{d}{d\beta_0}m(a,V \mid \beta_0)(Q_{02}(d(a),W) - m(a,V \mid \beta_0)),$$

which defines $\beta_0 = \beta(Q_{01}, Q_{02})$ as a function of the marginal distribution Q_{01} of W and the conditional distribution (i.e., mean) Q_{02}, of Y, given A,W. Let $\beta_n = \beta(Q_{1n}, Q_{2\theta_n}(\epsilon_n))$ be the targeted MLE, where Q_{1n} is the empirical probability distribution for the marginal distribution of W. It follows that, given Q_{1n} and $Q_{2\theta_n}(\epsilon_n)$, β_n can be defined as the solution of

$$0 = \frac{1}{n}\sum_{i=1}^n \sum_{a\in\mathcal{A}_1} h(a,V_i)\frac{d}{d\beta_n}m(a,V_i \mid \beta_n)(m(a,V \mid \beta_n) - Q_{2\theta_n}(\epsilon_n)(d(a),W_i))).$$

Equivalently, one can view β_n as a weighted least squares solution of the regression of $Q_{2\theta_n}(\epsilon_n)(d(a), W_i)$ on the realistic MSM $m(a, V_i \mid \beta)$:

$$\beta_n = \arg\min_{\beta} \sum_{a \in \mathcal{A}_1} h(a, V_i)(Q_{2\theta_n}(\epsilon_n)(d(a), W_i) - m(a, V_i \mid \beta))^2.$$

The targeted MLE as double robust estimating function based estimator: It is also important to note that

$$0 = \sum_{i=1}^{n} D_2^*(Q_{\theta_n}(\epsilon_n)) = 0,$$

so that

$$0 = \sum_i D^*(Q_{\theta_n}(\epsilon_n), g_n)(O_i).$$

Let's now use the estimating function representation of the efficient influence curve,

$$\begin{aligned} D^*(\beta, Q, g) &= \sum_{a \in \mathcal{A}_1} I(A = d(a)(W)) \frac{h(a, V) \frac{d}{d\beta} m(a, V \mid \beta)}{g(A \mid X)} (Y - Q_2(A, W)) \\ &\quad + \sum_{a \in \mathcal{A}_1} h(a, V) \frac{d}{d\beta} m(a, V \mid \beta)(Q_2(d(a), W) - m(a, V \mid \beta)), \end{aligned}$$

where $Q_2(a, W) = E_Q(Y \mid A = a, W)$ and $Q_2(d(a), W) = E_Q(Y \mid A = d(a)(W), W)$. We have $D^*(Q, g) = D^*(\beta(Q), Q, g)$ so that the fact that the targeted MLE $Q_{\theta_n}(\epsilon_n)$ solves the efficient influence curve equation, $P_n D^*(Q_{\theta_n}(\epsilon_n), g_n) = 0$, implies that the targeted MLE $\beta_n = \beta(Q_{\theta_n}(\epsilon_n))$ solves $P_n D^*(\beta_n, Q_{\theta_n}(\epsilon_n), g_n) = 0$. Thus the targeted MLE β_n is a solution of the double robust IPTW estimating function:

$$0 = \sum_i D^*(\beta_n, Q_{\theta_n}(\epsilon_n), g_n).$$

As a consequence, we can analyze β_n in the same manner as we analyze the double robust IPTW estimator β_{nDR} solving $0 = \sum_i D^*(\beta, Q_n, g_n)$ for a given estimator Q_n, but where Q_n is now simply playing the role of the updated $Q_{\theta_n}(\epsilon_n)$ (van der Laan, Robins, 2003).

Statistical Inference: Thus (van der Laan, Robins, 2003), if $g_n = g_0$, under regularity conditions, we have that the targeted MLE $\beta_n = \beta(Q_{1n}, Q_{2\theta_n}(\epsilon_n))$ is consistent and asymptotically linear with influence curve $c_0^{-1} D^*(\beta_0, Q^*, g_0)$,

where $c_0 = c(P_{Q_0,g_0}, g_0, \beta_0)$ is the derivative matrix defined above and Q^* denotes the limit of $Q_{\theta_n}(\epsilon_n)$ (which is allowed to be misspecified):

$$\beta_n - \beta_0 = \frac{1}{n}\sum_{i=1}^{n} c_0^{-1} D^*(Q_0, g_0)(O_i) + o_P(1/\sqrt{n}).$$

If the estimator g_n of g_0 is a maximum likelihood estimator for a correctly specified model for g_0, then this influence curve is known to be conservative and can thus still be used for conservative first order statistical inference. If one wants statistical inference in the double robust model only assuming that either g_n or $Q_{\theta_n}(\epsilon_n)$ is consistent, then we recommend to use the bootstrap.

2 Targeted MLE of causal effect of point treatment on survival.

Let $O = (W, A, T = T(A))$ be the observed data structure on an experimental unit, and let $X = (W, (T(a) : a))$ denote the full data on the unit consisting of the counterfactual survival times, and baseline covariates. Let $S_a(\cdot) = P(T_a > \cdot)$ be the treatment specific survival function. For a particular time point t we define an effect of treatment A on survival at time t as

$$\theta_0(t) \equiv f(S_1(t), S_0(t))$$

for some function $(x, y) \to f(x, y)$. Let $f_1(x, y) = \frac{d}{dx} f(x, y)$ and $f_0(x, y) = \frac{d}{dy} f(x, y)$ denote the two first order derivatives of f. The efficient influence curve for $\theta_0(t)$ at $P_{Q,g}$ in a nonparametric model, is thus given by

$$IC_t = f_1(S_1(t), S_0(t)) IC_{1t} + f_2(S_1(t), S_0(t)) IC_{0t},$$

where

$$\begin{aligned} IC_{1t} &= (Y(t) - Q(t)(A, W)) \frac{I(A = 1)}{g(1 \mid W)} + Q(t)(1, W) - S_1(t) \\ IC_{0t} &= (Y(t) - Q(t)(A, W)) \frac{I(A = 0)}{g(0 \mid W)} + Q(t)(0, W) - S_0(t). \end{aligned}$$

Here $g(a \mid W) = P(A = a \mid W)$ denotes the treatment mechanism, $Y(t) = I(T > t)$, and $Q(t)(A, W) = P(T > t \mid A, W)$.

For example, if $f_{COXPH}(S_1(t), S_0(t)) = \log S_1(t) / \log S_0(t)$, then we have

$$IC_{t,COX} = \frac{1}{S_1(t) \log S_0(t)} IC_{1t} - \frac{1}{S_0(t) \log^2 S_0(t)} IC_{0t}.$$

If $f_{AR}(S_1(t), S_0(t)) = S_1(t) - S_0(t)$ (AR stands for additive risk), we have $IC_{t,AR} = IC_{1t} - IC_{0t}$. If $f_{RR}(S_1(t), S_0(t)) = \log S_1(t)/S_0(t)$, then

$$IC_{t,RR} = \frac{1}{S_1(t)} IC_{1t} - \frac{1}{S_0(t)} IC_{0t}.$$

Finally, if $f_{OR}(S_1(t), S_0(t)) = \log S_1/(1-S_1)/S_0/(1-S_0)(t)$, then

$$IC_{t,OR} = \frac{1}{S_1(t)} IC_{1t} + \frac{1}{1-S_1(t)} IC_{1t} - \frac{1}{S_0(t)} IC_{0t} - \frac{1}{1-S_0(t)} IC_{0t}.$$

Let τ be a set of time-points or indices indexing such time points, such as a finite set of points within an interval $[a, b]$. Let $IC = (IC_t : t \in \tau)$ and $\Sigma = EIC(O)IC^\top(O)$ be the corresponding covariance matrix of this vector influence curve.

For simplicity, we focus here on a simple working model $\theta_0(t) = \gamma_0 = \exp(\beta_0)$, but our proposed class of tests of the null hypothesis $H_0 : S_1 = S_0$ generalize to general working models $\theta_0(t) = m(t \mid \beta_0)$ for some parametric model $m(t \mid \beta_0)$ indexed by possibly multivariate parameter β_0. In the latter case, our test statistic would be based on a test of $H_0 : \beta_0 = 0$, where the null value 0 is so that $m(t \mid 0)$ corresponds with the null hypothesis being true.

Let $\theta_n(t)$ be a targeted MLE of $\theta_0(t)$ in the nonparametric model (so not assuming the working model) based on data reduction $(W, A, Y(t) = I(T > t))$. Let Σ_n be an estimate of the covariance matrix Σ. Let

$$\gamma_n = \arg\min_\gamma (\theta_n - \gamma)^\top_{t\in\tau} \Sigma_n^{-1} (\theta_n - \gamma)_{t\in\tau}.$$

Let

$$\gamma_0 = \arg\min_\gamma (\theta_0 - \gamma)^\top_{t\in\tau} \Sigma^{-1} (\theta_0 - \gamma)_{t\in\tau}.$$

Under the assumption that the working model $m(t \mid \beta_0)$ (in this case $\exp(\beta_0)$) is correct, the first order asymptotics of the estimator γ_n is not affected by the estimating Σ so that we can just study the estimator γ_n with $\Sigma_n = \Sigma$, i.e. treating Σ as known. We wish to find the influence curve of γ_n as an estimator of γ_0.

For this purpose we note that

$$\gamma_n = g(\theta_n) \equiv \arg\min_\gamma \sum_{k,l\in\tau} \Sigma^{-1}(k,l)(\theta_n(k) - \gamma)(\theta_n(l) - \gamma).$$

Setting the derivative w.r.t γ equal to zero gives us the following equation for γ_n:

$$0 = \sum_{k,l\in\tau} \Sigma^{-1}(k,l)(\theta_n(k) + \theta_n(l) - 2\gamma).$$

This shows that

$$\gamma_n = g(\theta_n) \equiv \frac{\sum_{k,l\in\tau} \Sigma^{-1}(k,l)\frac{\theta_n(k)+\theta_n(l)}{2}}{\sum_{k,l\in\tau}\Sigma^{-1}(k,l)}.$$

This teaches us that (treating Σ as given)

$$\begin{aligned}
\gamma_n - \gamma_0 &= g(\theta_n) - g(\theta_0) \\
&= \frac{\sum_{k,l\in\tau} \Sigma^{-1}(k,l)\frac{\theta_n(k)-\theta_0(k)+\theta_n(l)-\theta_0(l)}{2}}{\sum_{k,l\in\tau}\Sigma^{-1}(k,l)} \\
&\approx \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{k,l\in\tau} \Sigma^{-1}(k,l)\frac{IC_k(O_i)+IC_l(O_i)}{2}}{\sum_{k,l\in\tau}\Sigma^{-1}(k,l)} \\
&\equiv \frac{1}{n}\sum_{i=1}^{n} IC_\gamma(O_i),
\end{aligned}$$

where IC_γ denotes the influence curve of γ_n as an estimator of γ_0.

Right censoring. If T is subject to right censoring, then we replace the TMLE $\theta_n(t)$ by the targeted MLE of $\theta_0(t)$, or of the bivariate parameter $(S_0(t), S_1(t))$ in the nonparametric model based on the data structure $(W, A, I(T \leq C), \min(T, C))$. This targeted MLE is presented elsewhere. An ad hoc inefficient solution would be to use the targeted MLE for the reduced data structure $(W, A, \Delta(t) = I(Y(t) \text{ observed}), \Delta(t)Y(t))$, treating $Y(t)$ as missing, or, equivalently, apply IPCW weights $I(\Delta(t) = 1)/\Pi_n(t, A, W)$ when estimating the regressions $Q(t)(A, W) = E(Y(t) \mid A, W)$, where $\Pi(t, A, W) = P(\Delta(t) = 1 \mid A, W)$ and Π_n is an estimator thereof. The influence curve is now the same as above but IC_{0t} and IC_{1t} are now modified by weighing the relevant components of IC_{0t} and IC_{1t} with $I(\Delta(t) = 1)/\Pi(t, A, W)$.

3 Targeted MLE of causal effect on mean count in nonparametric model based on Poisson regression.

Let $O = (W, A, Y = Y(A))$, and let the model for its distribution P_0 be nonparametric. Our goal is to estimate a real valued function of EY_1 and EY_0. We will compute the targeted MLE of the joint parameter (EY_1, EY_0)

and map it into a locally efficient targeted maximum likelihood estimator of any function $g(EY_1, EY_0)$ by simple substitution.

Recall that the efficient influence curve of EY_1 is $I(A = 1)/g(1 \mid W)(Y - Q(A, W)) + Q(1, W) - EY_1$. The component of this efficient influence curve correspnding with the tangent space of the conditional distribution of Y given A, W is thus $I(A = 1)/g(1)(Y - Q(A, W))$. Similarly, the component of the efficient influence curve of EY_0 corresponding with this tangent space of Y, given A, W, is $I(A = 0)/g(0)(Y - Q(A, W))$. To compute the targeted maximum likelihood estimator we need an initial estimator $P^0(Y|A, W)$ and then extend it with a parameter ϵ that has score equal to these two scores $(I(A = 1)/g(1 \mid W)(Y - Q(A, W)), I(A = 0)/g(0 \mid W)(Y - Q(A, W)))$. Given such an ϵ-fluctuation at the initial fit, we will compute the MLE of ϵ, and iterate the corresponding updating process for estimation of Y, given A, W, if needed. The resulting iterative targeted maximum likelihood estimator will be double robust locally efficient.

As a special case, we can choose as initial model a Poisson regression:

$$dP_0(Y|A, W) = Q_0(A, W)^Y exp(-Q_0(A, W))/Y!,$$

where $Q_0(A, W) = E_0(Y|A, W) = exp(h_0(A, W))$ for some function h_0. A particular model for h_0 is a linear model $h_0(A, W) = beta_0 + beta_1 A + beta_2 W$, but for now consider an arbitrary model. Let $Q^0(A, W) = exp(h^0(A, W))$ be an initial estimator of $Q_0(A, W)$ such as the MLE according to a Poisson regression model.

We now define an extension $P(Q^0(\epsilon))(Y|A, W)$ by substituting for $Q_0 = E(Y|A, W)$ $Q^0(\epsilon) = exp(h^0 + \epsilon C(A, W))$ with $C(A, W)$ a clever covariate specified below. The score of ϵ at $\epsilon = 0$ equals $C(A, W)(Y - Q^0(A, W))$. Thus, if we choose $C(A, W) = (I(A = 1)/g(1 \mid W), I(A = 0)/g(0 \mid W))$, then the score of ϵ at $\epsilon = 0$ equals the wished two efficient influence curve components.

Note that in a randomized trial adding the covariate $C(A, W)$ is equivalent to adding A. Let's consider this randomized trial application. The TMLE is obtained by adding to the initial h^0 these two covariates and computing the MLE of its coefficient ϵ, and iterate this process. However, if the initial h^0 is already a MLE and $h^0(A, W)$ contains a main term A, then in a randomized trial $C(A, W)$ is equivalent with the already present A. So if A is in model h^0 then the MLE of ϵ equals 0.

So if we do not estimate $g(A|W)$ in $C(A, W)$, but set it at its true value (say) 0.5, then the TMLE is just the original ML estimator. As a consequence, any function $g(EY_1, EY_0)$ can be locally efficiently estimated with

$g(E_nY_1, E_nY_0)$, where (E_nY_1, E_nY_0) is the MLE obtained by substituting Q^0. Thus, $E_nY_1 = 1/n \sum_i exp(h^0(1, W_i))$ and $E_nY_0 = 1/n \sum_i exp(h^0(0, W_i))$.

In particular, for the simple maximum likelihood model fit $h^0(A, W) = \beta_0 + \beta_1 A + \beta_2 W$, it follows that the TMLE of $g(EY_1, EY_0) = EY_1/EY_0$ equals $\exp(beta_1)$.

To conclude, in a randomized trial, the Poisson regression estimator of any $g(EY_1, EY_0)$ is locally efficient, since it equals the TMLE with initial density estimator the MLE according to Poisson regression in nonparametric model.

4 Obtaining special Robustness of Targeted MLE for parametric MSM in randomized trial.

Let $O = (W, A, Y = Y_A)$ and assume that A is independent of $X = (W, (Y_a : a))$, given $V \subset W$. Consider a marginal structural model $E(Y_a \mid V) = r(a\beta_1(1, V) + \beta_2(1, V))$ for some link function r. In general, we will consider models of the form

$$E(Y_a \mid V) = m_\beta(a, V) = r\left(\sum_{j=0}^{J} \beta_{1j} a^j h_j(V) + \sum_{j=0}^{J} \beta_{2j} g_j(V)\right),$$

where $g_j(V) = E(A^j \mid V)h_j(V))$. If A is completely randomized then $E(A^j \mid V)$ is just a constant so that these models reduce to $E(Y_a \mid V) = r(\sum_j \beta_{1j} a^j h_j(V) + \sum_j \beta_{2j} h_j(V))$.

In the previous section we defined a class of targeted MLE's for this parametric MSM, where we treated the parametric MSM as a working model so that it is used to define a causal parameter in a nonparametric model for the observed data distribution. Each choice of double robust estimating function indexed by a function $h(A, V)$ (equal to $h_1(A, V)\frac{d}{d\beta} m_\beta(A, V)$ for some h_1) corresponded with different nonparametric extensions (i.e., weighted least squares projections) of the causal parameter and thereby implied different targeted MLE's. That is, if $h = h_1 \frac{d}{d\beta} m_\beta$, then

$$\beta(Q) = \arg\min_\beta E_Q \sum_a h_1(a, V)(m_\beta(a, V) - E_Q(Y_a \mid V))^2,$$

which solves

$$0 = E_Q \sum_a h_1(a, V) \frac{d}{d\beta} m_\beta(a, V)(Q(a, W) - m_\beta(a, V)),$$

where $Q(a,W) = E_Q(Y \mid A = a, W)$. We showed that these h-specific targeted MLE could be implemented by adding to a linear or logistic regression fit Q^0 of Y, given A, W, an ϵ-extension $\epsilon h(A,V) g_0(A \mid V)/g_0(A \mid W)$, which in a randomized trial corresponds with $\epsilon h(A,V)$. If h does not depend on β itself, then the targeted MLE will converge in one step and, if the covariate $h(A,V)$ was already included in the initial logistic or linear regression fit, then the targeted MLE simply reduces to this initial fit.

We will consider two link functions of particular interest providing a surprising (i.e., more than asked for) robustness of a particular targeted MLE for this parametric MSM, corresponding with a particular choice of $h(A,V)$: the additive link function $r(x) = x$ and the multiplicative link function $r(x) = \exp(x)$. In addition, we will also study a marginal structural exponential regression model for a survival outcome which features the same special robustness of the targeted MLE.

The purpose of this section is to show that if $r(x) = x$ or $r(x) = \exp(x)$, then a particular choice of h results in a corresponding targeted MLE which is robust against misspecification of $E(Y_0 \mid V)$ in the sense that it yields a consistent and asymptotically linear estimator of β_{10} even if the V-component in the MSM is mis-specified. That is, if $r(x) = x$, then this targeted MLE (in the nonparametric model for this particular h-extension of the parametric MSM) yields a consistent and asymptotically linear estimator in the semi-parametric MSM $E(Y_a \mid V) - E(Y_0 \mid V) = \sum_j \beta_j a^j h_j(V)$, and if $r(x) = \exp(x)$, then the targeted MLE yields a consistent and asymptotically linear estimator in the semi-parametric MSM $E(Y_a \mid V) = r_0(V) \exp(\sum_j \beta_j a^j h_j(V))$, where r_0 is unspecified.

In order to establish these results we just need to show that the targeted MLE β_{1n} solves an unbiased estimating equation for the true β_{10}, even if the model components modelled by β_2-terms are misspecified. Given the efficient influence curve $D_h(Q, g_0)$ for the particular h-extension of the parametric MSM, we need to study the equation $E_0 D(Q, g_0) = 0$ and show that such a solution Q implies that $\beta_1(Q) = \beta_1(Q_0)$. We remind the reader that the targeted MLE Q_n of Q_0 solves $P_n D(Q_n, g) = 0$. For the sake of presentation we sometimes consider the simple linear parametric MSM and that A is completely randomized so that $E(A \mid V) = E(A)$, and in our theorems we present the general results. The efficient influence curve of the h-specific MSM parameter β can be represented as:

$$\begin{aligned} D_h(Q, g_0) &= h_\beta(A,V)(Y - Q(A,W)) \\ &\quad + \sum_a g_0(a \mid V) h_\beta(a,V)(Q(a,W) - r(a\beta_1(1,V) + \beta_2(1,V))), \end{aligned}$$

where $\beta = \beta(Q)$, h_β is a user supplied choice defining the non-parametric extension of β beyond the working model. That is, $h_\beta(A,V) = h_1(A,V)\frac{d}{d\beta}m(A,V \mid \beta)$, where h_1 can be viewed as a weight function, possibly also depending on β (e.g. involving inverse weighting by variance). Let $E_0(Y_a \mid V) = r(a\beta_{10}(1,V) + r_0(V))$. We have

$$\begin{aligned} E_0 D_h(Q,g_0) &= E_0 h_\beta(A,V)(Y - r(a\beta_1(1,V) + \beta_2(1,V)) \\ &= E_0 \sum_a g_0(a \mid V) h_\beta(a,V)(Y_a - r(a\beta_1(1,V) + \beta_2(1,V))) \\ &= E_0 \sum_a g_0(a \mid V) h_\beta(a,V)\{r(a\beta_{10}(1,V) + r_0(V)) - r(a\beta_1 V + \beta_2 V)\}. \end{aligned}$$

Thus, if Q solves $E_0 D_h(Q,g_0) = 0$, then this implies an equation for $\beta(Q)$, and we wish to establish that $\beta_1(Q) = \beta_{10} = \beta_1(Q_0)$. We will assume that the solution for β is unique.

4.1 Robustness of the targeted MLE for additive parametric MSM.

In this subsection we consider the case that $r(x) = x$, $h(a,V) = \frac{d}{d\beta} r_\beta(a,V)/\sigma^2(V) = (1,a,aV,V)/\sigma^2(V)$ for arbitrary given function $\sigma^2(V)$. A natural choice is $\sigma^2(V) = 1$. We note that this choice of h does not depend on β so that the targeted MLE will converge in a single step, and if the initial logistic or linear regression fit Q^0 already includes $(1,A,AV,V)$, then the targeted MLE reduces to the initial regression Q^0 of $E_0(Y \mid A,W)$.

Let's now consider a solution with $\beta_1 = \beta_{10}$, which gives the following equation for β_2:

$$\begin{aligned} 0 &= E_0 \sum_a g_0(a, \mid V) \frac{1}{\sigma^2(V)} (1,a,aV,V)^\top \{r_0(V) - \beta_2 V\} \\ &= E_0 \frac{1}{\sigma^2(V)} (1, E(A \mid V), E(A \mid V)V, V)^\top \{r_0(V) - \beta_2 V\}. \end{aligned}$$

Since $E(A \mid V)$ and $E(A \mid V)V$ are in the linear span of $(1,V)$, we have that the latter equation is solved by $\beta_{20} = \arg\min_{\beta_2} E_0 \frac{1}{\sigma^2(V)}(r_0(V) - \beta_2 V)^2$. Thus this shows that the unique solution $\beta(Q)$ satisfies $\beta_1(Q) = \beta_{10}$.

Theorem 1 *Let $O = (W, A, Y = Y_A)$ and assume that A is independent of $X = (W, (Y_a : a))$, given $V \subset W$. Let $g_0(A \mid X) = g_0(A \mid V)$ be the conditional probability distribution function of A, given X. Consider parametric marginal*

structural models of the form $E(Y_a \mid V) = m_\beta(a,V) = r(\sum_j \beta_{1j} a^j h_j(V) + \sum_j \beta_{2j} E(A^j \mid V) h_j(V))$ *for* $r(x) = x$. *If* A *is completely randomized then* $E(A^j \mid V)$ *is just a constant so that these models reduce to* $E(Y_a \mid V) = \sum_j \beta_{1j} a^j h_j(V) + \sum_j \beta_{2j} h_j(V))$. *The efficient influence curve of the h-specific nonparametrically defined MSM parameter* β *can be represented as:*

$$\begin{aligned} D_h(Q,g_0) &= h(A,V)(Y - Q(A,W)) \\ &\quad + \sum_a g_0(a \mid V) h(a,V)(Q(a,W) - \sum_j \beta_{1j} a^j h_j(V) - \sum_j \beta_{2j} E(A^j \mid V) h_j(V)), \end{aligned}$$

where $\beta = \beta(Q)$, *and*

$$h(a,V) = \frac{d}{d\beta} m_\beta(a,V)/\sigma^2(V)$$

for arbitrary given function $\sigma^2(V)$.

Let $E_{Q_0}(Y_a \mid V) = \sum_j \beta_{10j} a^j h_j(V) + r_0(V))$ *for some* $\beta_{10} = \beta_1(Q_0)$ *and* r_0. *We have that*

$$E_0 D(Q,g_0) = 0$$

is equivalent with

$$\begin{aligned} 0 &= E_0 \sum_a g_0(a \mid V) h(a,V) \{ \\ &\quad \sum_j \beta_{10j} a^j h_j(V) + r_0(V) - \sum_j \beta_{1j} a^j h_j(V) - \sum_j \beta_{2j} E(A^j \mid V) h_j(V)) \Big\}. \end{aligned}$$

Assume that the solution of this equation for β *is unique. Then any* Q *solving* $E_0 D(Q,g_0) = 0$ *satisfies* $\beta_1(Q) = \beta_{10} = \beta_1(Q_0)$.

4.2 Robustness of targeted MLE for multiplicative parametric MSM.

Now, we consider the case $r(x) = \exp(x)$. Recall that $E_0 D_h(Q,g_0) = 0$ is equivalent with

$$0 = E_0 \sum_a g_0(a \mid V) h_\beta(a,V) \left\{ \exp(a\beta_{10} V) \exp(r_0(V)) - - \exp(a\beta_1 V) \exp(\beta_2 V) \right\}.$$

We will now have to make the choice $h_\beta = \frac{d}{d\beta} m_\beta / m_\beta^2$ in order to achieve the wished robustness of the corresponding targeted MLE. Consider a solution

with $\beta_1 = \beta_{10}$, which gives the equation:

$$\begin{aligned} 0 &= E_0 \sum_a g_0(a \mid V) h_{\beta_{10},\beta_2}(a,V) \exp(\beta_{10} aV)\{\exp(r_0(V)) - \exp(\beta_2 V)\} \\ &= E_0 \sum_a g_0(a \mid V) \frac{1}{\exp(\beta_2 V)} (1, V, a, aV)^\top \{\exp(r_0(V) - \exp(\beta_2 V)\} \\ &= E_0 \frac{1}{\exp(\beta_2 V)} (1, V, E(A \mid V), E(A \mid V)V)\{\exp(r_0(V) - \exp(\beta_2 V)\}. \end{aligned}$$

Consider now the solution $\beta_{20} = \arg\min_{\beta_2} E_0 \frac{1}{\exp^2(\beta_2 V)} \{\exp r_0(V) - \exp(\beta_2 V)\}^2$, which solves the wished equation since $E(A \mid V) = E(A)$. Thus this shows that the unique solution $\beta(Q)$ satisfies $\beta_1(Q) = \beta_{10}$.

Theorem 2 *Let $O = (W, A, Y = Y_A)$ and assume that A is independent of $X = (W, (Y_a : a))$, given $V \subset W$. Let $g_0(A \mid X) = g_0(A \mid V)$ be the conditional probability distribution function of A, given X. Consider parametric marginal structural models of the form $E(Y_a \mid V) = m_\beta(a, V) = r(\sum_j \beta_{1j} a^j h_j(V) + \sum_j \beta_{2j} E(A^j \mid V) h_j(V))$ for $r(x) = \exp(x)$. If A is completely randomized, then $E(A^j \mid V)$ is just a constant so that these models reduce to $\log(E(Y_a \mid V)/E(Y_0 \mid V)) = \sum_j \beta_{1j} a^j h_j(V) + \sum_j \beta_{2j} h_j(V)$.*

The efficient influence curve of the h-specific nonparametric extended parametric MSM parameter β can be represented as:

$$\begin{aligned} D_h(Q, g_0) &= h_\beta(A, V)(Y - Q(A, W)) \\ &\quad + \sum_a g_0(a \mid V) h_\beta(a, V) \left(Q(a, W) - \exp\left(\sum_j \beta_{1j} a^j h_j(V) - \sum_j \beta_{2j} g_j(V) \right) \right), \end{aligned}$$

where $g_j(V) = E(A^j \mid V) h_j(V)$, $\beta = \beta(Q)$, and h_β is a user supplied choice defining the non-parametric extension of β beyond the working model. That is, if $h_\beta = h_{1\beta} \frac{d}{d\beta} m_\beta$, then

$$\beta(Q) = \arg\min_\beta E_Q \sum_a h_{1\beta(Q)}(a, V)(m_\beta(a, V) - E_Q(Y_a \mid V))^2.$$

which solves

$$0 = E_Q \sum_a h_{1\beta(Q)}(a, V) \frac{d}{d\beta} m_\beta(a, V)(Q(a, W) - m_\beta(a, V)),$$

where $Q(a,W) = E_Q(Y \mid A=a,W)$. *We set*

$$\begin{aligned} h_\beta(a,V) &= \frac{\frac{d}{d\beta} m_\beta(a,V)}{m_\beta^2(a,V)} \\ &= ((a^j h_j(V) : j), (E(A^j \mid V) h_j(V) : j))^\top / m_\beta(a,V) \\ &\equiv h*(a,V)/m_\beta(a,V). \end{aligned}$$

Let $E_{Q_0}(Y_a \mid V) = \exp\left(\sum_j \beta_{10j} a^j h_j(V) + r_0(V)\right)$ *for some* $\beta_{10} = \beta_1(Q_0)$ *and* r_0. *We have that*

$$E_0 D(Q, g_0) = 0$$

is equivalent with

$$\begin{aligned} 0 &= E_0 \sum_a g_0(a \mid V) h_\beta(a,V) \{ \\ &\exp\left\{\sum_j \beta_{10j} a^j h_j(V) + r_0(V)\right\} - \exp\left\{\sum_j \beta_{1j} a^j h_j(V) + \sum_j \beta_{2j} E(A^j \mid V) h_j(V)\right\}\Bigg\}. \end{aligned}$$

Assume that the solution of this equation for β *is unique. Then any* Q *satisfying* $E_0 D(Q, g_0) = 0$ *satisfies* $\beta_1(Q) = \beta_{10} (= \beta_1(Q_0))$ *and* $\beta_2(Q)$ *is the solution of*

$$0 = E_0 \frac{1}{\exp\left(\sum_j \beta_{2j} g_j(V)\right)} (g_j(V) : j)^\top \{\exp(r_0(V) - \exp(\beta_2 V)\}.$$

Implementation of this h-specific targeted MLE for the parametric MSM requires now adding a covariate $h_{\beta_n^0}(A,V) = \frac{d}{d\beta_n^0} m_{\beta_n^0} / m^2_{\beta_n^0}$ to (e.g.) an initial logistic regression fit, where this covariate depends on β. As a consequence, the computation of this targeted MLE requires now iteration in order to solve the efficient influence curve equation $P_n D(Q_n, g_0) = 0$. Normally, one might implement the targeted MLE by adding the covariate $h = \frac{d}{d\beta} m_\beta / m_\beta$ to a logistic regression fit, where this covariate h does not depend on β so that the targeted MLE converges in a single update step (where the single update step might not even be necessary since the current regression fit Q_n^0 might already contain these covariates). Apparently, among the class of targeted MLE for the parametric multiplicative MSM, indexed by different choices h, the choice which happens to give the surprising robustness corresponds with adding a special covariate h_β depending on β so that iteration of the targeted MLE update step is now necessary.

4.3 Robustness of targeted MLE for parametric marginal structural exponential regression model.

The robustness of the targeted MLE for the parametric MSM $P(Y_a = 1 \mid V) = \exp(a(\beta_1 V))\exp(\beta_2 V)$ w.r.t. miss-specification of $P(Y_0 = 1 \mid V)$, implying the result that the targeted MLE provides a locally efficient estimator of the relative risk $\exp(a\beta_{10}V)$ in the semi-parametric MSM $P(Y_a = 1 \mid V)/P(Y_0 = 1 \mid V) = \exp(a\beta_1 V)$, motivates us to see if we can extend this result to marginal structural exponential hazard regression models.

Suppose we observe $O = (W, A, (Y_A(t) : t))$, where W are baseline covariates, A is treatment, and $Y_a(t)$ is a time-dependent counting process. We assume that A is independent of $X = (W, (Y_a : a))$, given V, and let $g_0(\cdot \mid X) = g_0(\cdot \mid V)$ be the conditional probability distribution of A, given V.

Consider now the model

$$\begin{aligned} E(dY_a(t) \mid \bar{Y}_a(t-), V) &= Y_a^*(t)\lambda_\beta(a, V) \\ &\equiv Y_a^*(t)\exp(\sum_{j=0}^{J}\beta_{1j}a^j h_j(V) + \sum_{j=0}^{J}\beta_{2j}g_j(V)), \end{aligned}$$

where $g_j(V) \equiv E(A^j \mid V)h_j(V)$ and $Y_a^*(t)$ is an indicator of $Y_a(t)$ being at risk of jumping at time t, which is a function of $\bar{Y}_a(t-), V$. In other words, if $Y_a(t) = I(T_a \leq t)$, then this model assumes an exponential distribution for T_a, given V, with mean given by $\lambda_\beta(a, V)$. An example is the simple linear parametric MSM

$$E(dY_a(t) \mid \bar{Y}_a(t-), V) = Y_a^*(t)\lambda_\beta(a, V) = Y_a^*(t)\exp\left(a\beta_1(1, V) + \beta_2(1, V)\right),$$

with β_1 and β_2 two dimensional parameters,

The purpose of this section is to show that the targeted MLE is robust against miss-specification of $E(dY_0(t) \mid \bar{Y}_0(t-), V)$ in the sense that it yields a consistent and asymptotically linear estimator of β_{10} even if the V-component $\sum_j \beta_{2j}g_j(V)$ in the MSM is miss-specified.

In order to establish this result we need to show that the targeted MLE β_{1n} solves an unbiased estimating equation for the true β_{10}, even if the model components modelled by β_2-terms is misspecified. Given the efficient influence curve $D(Q, g_0)$ for the parametric MSM the targeted MLE is based upon, we need to study the equation $E_0 D(Q, g_0) = 0$ and show that such a solution Q implies that $\beta_1(Q) = \beta_1(Q_0)$: recall that the targeted MLE Q_n of Q_0 solves $P_n D(Q_n, g) = 0$. We represent $\lambda_\beta(a, V) = \lambda_{1\beta_1}(a, V)\lambda_{2\beta_2}(V)$. The efficient

influence curve of the MSM parameter β can be represented as:

$$\begin{aligned} D(Q,g_0) &= \sum_t h_\beta(A,V)\{dY_A(t) - Y_A^*(t)\lambda_\beta(A,V)\} \\ &\quad - \sum_t h_\beta(A,V)\{q(t,A,W) - \bar{Q}(t,A,W)\lambda_\beta(A,V)\} \\ &\quad + \sum_t \sum_a g_0(a \mid V) h_\beta(a,V)\{q(t,a,W) - \bar{Q}(t,a,W)\lambda_\beta(a,V)\}, \end{aligned}$$

where $\beta = \beta(Q)$, $q(t,A,W) = E(dY(t) \mid A,W)$ and $\bar{Q}(t,A,W) = E(Y^*(t) \mid A,W)$, and h_β is a user supplied choice defining the non-parametric extension of β beyond the working model (defined as solution of expectation of last term). We will choose $h_\beta(A,V) = \frac{d}{d\beta}\lambda_\beta(A,V)/\lambda_\beta(A,V)$. Let the true causal hazard be given $\lambda_0(a,V) = \exp(a\beta_{10}(1,V) + r_0(V)) = \lambda_{01}(a,V)\lambda_{02}(V)$ for some function $r_0(V)$, and let $\bar{Q}_0^*(a,t,V) = E_0(Y_a^*(t) \mid V)$. We have

$$\begin{aligned} E_0 D(Q,g_0) &= E_0 \sum_t h_\beta(A,V)(dY_A(t) - Y_A^*(t)\lambda_\beta(t,A,V)) \\ &= E_0 \sum_t \sum_a g_0(a \mid V) h_\beta(a,V)(dY_a(t) - Y_a^*(t)\lambda_\beta(t,a,V)) \\ &= E_0 \sum_t \sum_a g_0(a \mid V) h(a,V) Y_a^*(t)(\lambda_0(a,V) - \lambda_\beta(a,V)) \\ &= E_0 \sum_t \sum_a g_0(a \mid V) h(a,V) \bar{Q}_0^*(a,t,V)(\lambda_0(a,V) - \lambda_\beta(a,V)) \\ &= E_0 \sum_t \sum_a g_0(a \mid V) h(a,V) \bar{Q}_0^*(a,t,V)\{\lambda_{10}(a,V)\lambda_{20}(V) - \lambda_{1\beta_1}(a,V)\lambda_{2\beta_2}(V)\} \\ &= E_0 \sum_t \sum_a g_0(a \mid V) h(a,V) \bar{Q}_0^*(a,t,V)\lambda_{2\beta_2}(V)\{\lambda_{10}(a,V) - \lambda_{1\beta_1}(a,V)\} \\ &\quad + E_0 \sum_t \sum_a g_0(a \mid V) h(a,V) \bar{Q}_0^*(a,t,V)\lambda_{10}(a,V)\{\lambda_{20}(V) - \lambda_{2\beta_2}(V)\}. \end{aligned}$$

We assume that this equation in β has a unique solution. As a candidate solution we consider a solution with $\beta_1 = \beta_{10}$. Note that this choice makes the first term equal to zero. We now consider the second term. Firstly, we assume that $\bar{Q}_0^*(a,t,V)\lambda_0(t,a,V)$ (e.g., if $Y_a(t) = I(T_a \le t)$, then it equals $P(T_a = t \mid V)$) is the conditional density in t so that

$$\sum_t \bar{Q}_0^*(a,t,V)\lambda_{10}(a,V) = \frac{\sum_t \bar{Q}_0^*(a,t,V)\lambda_0(t,a,V)}{\lambda_{20}(V)} = 1/\lambda_{02}(V).$$

Note that $h_\beta(a,V)=(h_1(a,V),h_2(V))$. So the second term corresponds with the following two equations for β_2:

$$E_0E_0(h_1(A,V)\mid V)\frac{1}{\lambda_{20}(V)}\left\{\lambda_{20}(V)-\lambda_{2\beta_2}(V)\right\}$$
$$E_0h_2(V)\frac{1}{\lambda_{20}(V)}\left\{\lambda_{20}(V)-\lambda_{2\beta_2}(V)\right\}.$$

Now, note that our choice satisfies:

$$h_\beta(A,V)=\frac{\frac{d}{d\beta}\lambda_\beta(A,V)}{\lambda_\beta(A,V)}=((A^jh_j(V):j),(E(A^j\mid V)h_j(V):j)),$$

where $h_1(A,V)=(A^jh_j(V):j)$ and $h_2(V)=(E_0(A^j\mid V)h_j(V):j)$. So for this choice we have $E_0(h_1(A,V)\mid V)=h_2(V)$, which proves that $\beta_{20}=\arg\max_{\beta_2}E_0h_2(V)/\lambda_{20}(V)\left\{\lambda_{20}-\lambda_{2\beta_2}\right\}^2(V)$ solves both equations.

This proves the following theorem.

Theorem 3 *Consider the model*

$$\begin{aligned}E(dY_a(t)\mid\bar{Y}_a(t-),V) &= Y_a^*(t)\lambda_\beta(a,V)\\ &\equiv Y_a^*(t)\exp(\sum_{j=0}^J\beta_{1j}a^jh_j(V)+\sum_{j=0}^J\beta_{2j}g_j(V)),\end{aligned}$$

where $g_j(V)\equiv E(A^j\mid V)h_j(V)$ *and* $Y_a^*(t)=I(T_a\geq t)$ *is an indicator of* $Y_a(t)=I(T_a\leq t)$ *being at risk of jumping at time* t.

Consider the efficient influence curve of the MSM parameter β

$$\begin{aligned}D(Q,g_0) &= \sum_t h_\beta(A,V)\{dY_A(t)-Y_A^*(t)\lambda_\beta(A,V)\}\\ &\quad-\sum_t h_\beta(A,V)\{q(t,A,W)-\bar{Q}(t,A,W)\lambda_\beta(A,V)\}\\ &\quad+\sum_t\sum_a g_0(a\mid V)h_\beta(a,V)\{q(t,a,W)-\bar{Q}(t,a,W)\lambda_\beta(a,V)\},\end{aligned}$$

where $\beta=\beta(Q)$, $q(t,A,W)=E(dY(t)\mid A,W)$ *and* $\bar{Q}(t,A,W)=E(Y^*(t)\mid A,W)$, *and*

$$h_\beta(A,V)=\frac{d}{d\beta}\lambda_\beta(A,V)/\lambda_\beta(A,V)=((A^jh_j(V):j),(E(A^j\mid V)h_j(V):j)).$$

Let the true causal hazard be given by $\lambda_0(a,V) = \exp(a\beta_{10}(1,V) + r_0(V)) = \lambda_{01}(a,V)\lambda_{02}(V)$ *for some function* $r_0(V)$*, represent* $\lambda_\beta = \lambda_{1\beta_1}\lambda_{2\beta_2}$*, and let* $\bar{Q}_0^*(a,t,V) = E_0(Y_a^*(t) \mid V)$*. We have*

$$E_0 D(Q, g_0) = 0$$

implies

$$\begin{aligned} 0 &= E_0 \sum_a g_0(a \mid V) h(a,V) \frac{1}{\lambda_0(a,V)} \lambda_{2\beta_2}(V) \left\{\lambda_{10}(a,V) - \lambda_{1\beta_1}(a,V)\right\} \\ &\quad + E_0 \sum_a g_0(a \mid V) h(a,V) \frac{1}{\lambda_0(a,V)} \lambda_{10}(a,V) \left\{\lambda_{20}(V) - \lambda_{2\beta_2}(V)\right\}. \end{aligned}$$

We assume that this equation in β *has not more than one solution. Then any* Q *solving* $E_0 D(Q, g_0) = 0$ *satisfies* $\beta_1(Q) = \beta_1(Q_0) = \beta_{10}$*, and* $\beta_2(Q) = \arg\max_{\beta_2} E_0 h_2(V)/\lambda_{20}(V) \left\{\lambda_{20} - \lambda_{2\beta_2}\right\}^2 (V)$.

5 Relation between efficiency of targeted MLE of causal effect in randomized trial and the prediction performance of initial regression estimator.

The approach in this section can be applied in general. Let

$$D_{1Q}(W,A,Y) = \frac{I(A=1)}{g_0(1)}(Y - Q(1,W)) + Q(1,W)$$

be the influence curve of the targeted MLE of $EY_1 = E_0 E_0(Y \mid A = 1, W)$ based on i.i.d. sampling from $p_0 = Q_0 g_0$, using the true known treatment mechanism g_0, and an initial regression estimator Q_n^0 of $E_0(Y \mid A = 1, W)$ converging to a Q as $n \to \infty$. The variance of this influence curve D_Q under $P_0 = Q_0 g_0$ as a function of Q is minimized at the true $Q = Q_0(1,W) = E_0(Y \mid A = 1, W)$, which shows that the targeted MLE of the treatment specific mean EY_1 in a randomized trial is most efficient if one correctly estimates the true $Q_0(A,W)$. In this section, we investigate how the variance of D_{1Q} (and thus also D_{0Q}) under P_0 depends on Q so that we can determine in what sense Q should approximate Q_0. This will teach us that to obtain a maximally efficient targeted ML estimator of the causal effect in randomized trials one should use machine learning algorithms combined with an aggressive

use of cross-validation as in super learning (**?**) to obtain the initial regression estimator.

By Theorem 1 in Rubin and VanderLaan (2007), we have

$$\begin{aligned} \mathrm{VAR}_{P_0} D_{1Q}(O) &= C_0 + E_0 \frac{1-g_0(1)}{g_0(1)}(Y_1 - Q(1,W))^2 \quad (2) \\ &= C_0 + E_0 I(A=1)\frac{1-g_0(1)}{g_0(1)^2}(Y - Q(A,W))^2, \quad (3) \end{aligned}$$

where C_0 does not depend on Q. Equivalently,

$$\mathrm{VAR}_{P_0} D_{1Q}(O) = C_{10} + E_0 \frac{1-g_0(1)}{g_0(1)}(Q_0(1,W) - Q(1,W))^2,$$

where C_{10} does not depend on Q.

Thus, the asymptotic variance of the targeted MLE $\psi_n(Q_n) = \frac{1}{n}\sum_i Q_n(1,W_i)$ of EY_1 equals a constant (only having to do with the data generating distribution but not with the choice Q_n) plus the asymptotic squared error loss risk $E_0 \frac{I(A=1)(1-g_0(1))}{g_0(1)^2}(Y - Q(A,W))^2$ of the asymptotic limit of Q_n. This proves that any improvement in the squared error risk of Q_n on observations with $A = 1$ immediately translates (in a linear manner) in a gain in variance for the resulting estimator EY_1. To be specific, let's consider a randomized trial with $g_0(1) = 0.5$. In this case, we have that the asymptotic variance of the targeted MLE of EY_1 equals a constant plus $2 * E_0 I(A=1)(Y - Q(1,W))^2$. In particular, given limits Q_1 and Q_2 corresponding with two targeted MLE Q_{1n} and Q_{2n}, and a limit $Q^* = E(Y \mid A)$ of the targeted MLE ignoring any covariates (and thereby corresponds with the unadjusted estimator $\sum_i Y_i I(A_i = 1)/\sum_i I(A_i = 1)$), we have the relation

$$\frac{VAR(D_{Q_1}) - VAR(D_{Q^*})}{VAR(D_{Q_2}) - VAR(D_{Q^*})} = \frac{E_0 I(A=1)(Y - Q^*(1,W))^2 - E_0 I(A=1)(Y - Q_1(1,W))^2}{E_0 I(A=1)(Y - Q^*(1,W))^2 - E_0 I(A=1)(Y - Q_2(1,W))^2}.$$

We can also relate the relative efficiency of the targeted MLE indexed by Q_1 relative to the targeted MLE indexed by Q^* in terms of the gain in risk w.r.t to squared error loss function defined as

$$RD_1 \equiv E_0 I(A=1)(Y - Q^*(1,W))^2 - E_0 I(A=1)(Y - Q_1(1,W))^2$$

and the variance $VAR(D_{Q^*})$. That is,

$$\mathrm{RE}_1 \equiv \frac{VARD_{1Q_1}}{VARD_{1Q^*}} = 1 + \frac{RD_1}{VARD_{1Q^*}}.$$

Thus, a gain in prediction performance as measured by cross-validated risk of the squared error loss function on the sample with $A = 1$ (i.e., RD_1), and an estimate of the standardized variance of the naive targeted MLE indexed by Q^* (i.e., $VARD_{1Q^*}$) maps into an estimate of the relative efficiency.

6 General strategy for constructing hardest ϵ-submodels to define targeted MLE update.

Here we present a useful strategy for constructing a ϵ-fluctuation through an initial fit P^0 of the probability distribution P_0 of O in model $\mathcal{M}$ with score at $\epsilon = 0$ equal to the efficient influence curve. Consider a model $\mathcal{M}$ and an initial model based estimate $P^0 \in \mathcal{M}$ of the true distribution P_0 of O. Consider a class of ϵ-extensions $\{P_h^0(\epsilon) : \epsilon\} \subset \mathcal{M}$ indexed by directions h. In most of the applications presented here this corresponds with adding a covariate h to an initial regression fit. As a next step, one computes the corresponding set of scores $\mathcal{S}(P^0) = \left\{S(h) = \frac{d}{d\epsilon} \log dP_h^0(\epsilon)/dP^0\big|_{\epsilon=0} : h\right\}$ of each of these h-specific ϵ-fluctuations of the initial fit P^0 of the true probability distribution P_0. Let $T_{nuis}^{\perp}(P^0)$ be a subset of the, or the whole, orthogonal complement of the nuisance tangent space at P^0, as defined in van der Laan, Robins (2003) and presented for many models, parameters of interest, and observed data structures. Now, find an h^* so that $S(h) \in T_{nuis}^{\perp}(P^0)$. Then $S(h)$ is in the linear span of the components of the efficient influence curve, so that this choice h^* identifies a wished hardest sub-model $\{P_{h^*}^0(\epsilon) : \epsilon\}$ through P^0. We illustrate this approach in the following two applications for additive and multiplicative variable importance.

7 The targeted MLE for model based additive variable importance.

Let $O = (W, A, Y)$, A is the variable whose variable importance we target, and Y is an outcome of interest. Assume the model

$$E_0(Y \mid A, W) - E_0(Y \mid A = 0, W) = m(A, W \mid \beta_0).$$

We wish to estimate β_0 with a targeted MLE. In many applications the data on one subject is a list of variables such as biomarkers, single nucleotide polymorphisms (SNP), gene expressions and so on, and an outcome of interest. This additive variable importance parameter $m(A, W \mid \beta_0)$ represents the effect of one of these variables A on outcome Y adjusting for a set W. Simple

models one might choose are linear models such as $m(A,W \mid \beta_0) = \beta_0 A$ or $m(A,W \mid \beta_0) = A(\beta_0 W)$. By carrying out the targeted MLE estimate of β_0 for each definition of a variable A and corresponding set W, separately, one obtains a list of targeted MLE estimates of each variable importance and corresponding p-values and standard error estimates. This has important applications in biomarker discovery and effect modification analysis. One important application is that A represents a randomized treatment in a clinical trial and W is set equal to a particular small subset W_j of a large list of biomarkers/genomic/genetic markers, where one can carry out this analysis for a large set of possibly subsets W_j, $j = 1, \ldots, J$.

In (Tuglus, van der Laan, 2007) we compare the practical performance (based on simulated data) of this targeted MLE method for estimating the above measure of variable importance and for obtaining a corresponding ranking of the variables by their importance with the current methods for ranking variables based on univariate linear regression and random forest.

We now proceed with deriving the targeted MLE using the method outlined in the previous section, as is also presented in van der Laan, Rubin (2006). The orthogonal complement of nuisance tangent space at P_0 contains the class of functions $\{h_0(A \mid W)(Y - m(A,W \mid \beta_0) - E(Y \mid A=0,W) : E(h_0(A \mid W) \mid W) = 0\}$, and the optimal index h_0^* corresponding with efficient influence curve is such that

$$h_0^*(A \mid W) = \left\{ \frac{d}{d\beta^0} m(A,W \mid \beta_0) - \frac{E(\frac{d}{d\beta_0} m(A,W \mid \beta_0)/\sigma^2(A,W) \mid W)}{E(1/\sigma^2(A,W) \mid W)} \right\} \frac{1}{\sigma^2(A,W)}.$$

If the conditional variance of Y, given A, W, only depends on W, i.e., $\sigma^2(A,W) = \sigma^2(W)$, then this optimal index simplifies to

$$h_0^*(A \mid W) = \{ \frac{d}{d\beta_0} m(A,W \mid \beta_0) - E(\frac{d}{d\beta_0} m(A,W \mid \beta_0) \mid W) \} \frac{1}{\sigma^2(W)}.$$

Consider the following ϵ-extension of an initial model based fit $Q^0(A,W) = m(A,W \mid \beta^0) + r^0(W)$ of $E(Y \mid A,W)$:

$$Q^0(\epsilon)(A,W) = m(A,W \mid \beta^0 + \epsilon) + r^0(W) + \epsilon r(W).$$

Let $Q^0(\epsilon)(Y \mid A,W)$ be a normal distribution with mean $Q^0(\epsilon)(A,W)$ and variance $\sigma^2(A,W)$. We have that

$$\begin{aligned} S(r) &= \frac{d}{d\epsilon} \log Q^0(\epsilon)(Y \mid A,W) \Big|_{\epsilon=0} \\ &= \left\{ \frac{d}{d\beta^0} m(A,W \mid \beta^0) + r(W) \right\} \frac{1}{\sigma^2(A,W)} (Y - Q^0(A,W)). \end{aligned}$$

In order to make the right-hand side equal to a score in the orthogonal complement of the nuisance tangent space at Q^0, and thereby equal to the efficient influence curve up till a normalizing matrix, where we remind the reader that the latter orthogonal complement contains the functions $\{(h(A,W)-E(h(A,W) \mid W))(Y-Q^0(A,W)) : h\}$, we need that

$$r(Q^0,g)(W) = -\frac{E_g(\frac{d}{d\beta^0}m(A,W \mid \beta^0)/\sigma^2(A,W) \mid W)}{E_g(1/\sigma^2(A,W) \mid W)}.$$

If $\sigma^2(A,W)$ only depends on W, then this simplifies to

$$r(Q^0,g)(W) = -E_g(\frac{d}{d\beta^0}m(A,W \mid \beta^0) \mid W).$$

This defines now our wished ϵ-fluctuation through an initial estimate $Q^0(Y \mid A,W)$ of the conditional distribution of Y, given A,W, so that we are ready to define the targeted MLE update.

It follows that the first-step targeted MLE of β_0 is given by $\beta_n = \beta_n^0 + \epsilon_n$, where

$$\epsilon_n = \arg\min_{\epsilon} \sum_i \frac{1}{\sigma^2(A_i,W_i)}(Y_i - m(A_i,W_i \mid \beta_n^0+\epsilon) + \epsilon r(Q^0,g_n)(W_i))^2.$$

If $m(\cdot \mid \beta)$ is linear in β, say $m(A,W \mid \beta) = A\beta W$, then $\frac{d}{d\beta}m(A,W \mid \beta) = AW$ so that (e.g., if $\sigma^2(A,W)=\sigma^2(W)$), $r(Q^0,g)(W) = r(g)(W) = WE_g(A \mid W)$, so that $r(g_n)$ only involves estimating the regression of A on W. In this case, ϵ_n exists in closed form as a linear regression least squares estimator.

If $\frac{d}{d\beta}m(A,W \mid \beta)$ does not depend on β (i.e., m is linear in β) so that $r(Q^0,g) = r(g)$ does not depend on β, then this first step targeted MLE is the targeted MLE, since iteration of this update step will not result in further changes (note that g does not get updated in this targeted MLE step).

Statistical Inference: We have that β_n solves the double robust estimating equation:

$$0 = \sum_i h_n(A_i,W_i)(Y_i - m(A_i,W_i \mid \beta_n) - r(Q^0,g_n)(\epsilon_n)(W_i)),$$

where

$$h(Q^0,g_n)(A,W) = \left\{\frac{d}{d\beta^0}m(A,W \mid \beta^0) + r(Q^0,g_n)(W)\right\}\frac{1}{\sigma^2(A,W)}.$$

Statistical inference can now be based on the influence curve of this estimating equation in β (i.e., the estimating function itself standardized by minus the

inverse of the derivative matrix), where one relies on correct specification of the regression of A on W. If one wished to rely on the double robustness, then one could use the bootstrap.

8 Delta-additive variable importance.

For biomarker discovery it is important to detect the variables with a causal effect on the outcome. This requires selecting the adjustment set W as large as possible so that all measured confounders of the effect of A on outcome Y are included. As the simulations in Tuglus, van der Laan (2007) show, this is the reason that our measure of variable importance outperforms univariate regression. On the other hand, if a variable in the adjustment set W has a very high correlation (e.g., 0.9) with the current variable A of interest, then aiming to adjust for such a variable can hurt the performance of the variable importance estimate and thereby deteriorate the ranking by variable importance for a list of biomarkers. One scenario which helps to explain this phenomena is the following. Suppose that one variable in the adjustment set W has an extreme correlation of say 0.999 with A. Without guidance, the targeted MLE for the additive variable importance of A will now aim to adjust for this perfectly correlated confounder, and thereby will not allow any adjustment by other confounders. Since it is impossible to disentangle the effect of this perfect confounder in W from the effect of A, the targeted MLE is aiming to do an impossible job, and thereby fails to succeed in doing the possible jobs of adjusting for the other potential confounders. Therefore, we have proposed to compute a δ-W-adjusted variable importance, which only adjusts for all confounders in W which have a correlation smaller than δ with A, and we compute this for each value of δ ranging from 0 to 1. In this manner, one obtains a whole curve of variable importance measures ranging from the unadjusted univariate regression ($\delta = 0$) to the fully adjusted variable importance adjusting for the complete set of confounders W. In combination of knowing the adjustment set $W(\delta)$ for each value of δ, this sequence of variable importance measures and corresponding p-values and confidence intervals provides important information and a complete picture. In particular, one can obtain a data adaptive recommendation of the choice δ for the purpose of obtaining an accurate estimate of the fully adjusted variable importance, by aiming to minimize a mean squared error over δ.

9 Targeted MLE for model based Multiplicative variable importance.

In the previous section the effect of interest was measured on the additive scale. In some applications people prefer to measure the effect of a variable on a multiplicative scale. Let $O = (W, A, Y)$, where A is the variable whose variable importance we target, and the outcome Y could be a count or have only two outcomes $\{0, 1\}$. Assume

$$\log \frac{E(Y \mid A, W)}{E_0(Y \mid A = 0, W)} = m(A, W \mid \beta_0).$$

Equivalently,

$$\frac{E_0(Y \mid A = 0, W)}{E_0(Y \mid A, W)} = m^*(A, W \mid \beta_0) \equiv \exp(-m(A, W \mid \beta_0)).$$

We wish to construct a targeted MLE of the unknown parameter β_0 and thereby of the importance $\exp(-m(A, W \mid \beta_0))$of variable A in predicting the outcome Y.

The orthogonal complement of the nuisance tangent space contains

$$\{h(A \mid W)(Ym^*(A, W) - E_0(Y \mid A = 0, W)) : E_0(h(A \mid W) \mid W) = 0\},$$

and one choice $h^*(A \mid W)$ results in the efficient influence curve. This subset of the orthogonal complement of the nuisance tangent space can also be represented as:

$$\{m^*(A, W \mid \beta_0)h(A \mid W)(Y - E(Y \mid A, W)) : E(h(A \mid W) \mid W) = 0\}. \quad (4)$$

Deriving the wished ϵ-extension: Consider an initial model based fit $Q^0(A, W)$ of $E(Y \mid A, W)$: $\log Q^0(A, W) = m(A, W \mid \beta^0) + r^0(W)$. Consider the class of ϵ-extensions $Q^0(\epsilon) \log Q^0(\epsilon)(A, W) = m(A, W \mid \beta^0 + \epsilon) + r^0(W) + \epsilon r(W)$ indexed by an arbitrary choice r. We follow the general strategy described above to determine the hardest sub-model whose score at $\epsilon = 0$ is in the linear span of the efficient influence curve of β_0. We will do this for the case that Y is binary. The same strategy can be worked out for Y being discrete where (e.g.,) we assume that $Q^0(Y \mid A, W)$ and $Q^0(\epsilon)(Y \mid A, W)$ follows a Poisson distribution, and the ϵ-extension involves modifying the mean of the Poisson distribution and it is selected so that the score of this Poisson likelihood in ϵ at $\epsilon = 0$ is an element of $T^{\perp}_{nuis}(P^0)$. The latter we will do at the end of this section.

The ϵ-extension for Bernoulli outcomes: We first wish to calculate the score of this extension at $\epsilon = 0$. This score is given by

$$S(r) = \frac{1}{1 - Q^0(A,W)}\{\frac{d}{d\beta} m(A,W \mid \beta^0) + r(W)\}(Y - Q^0(A,W)).$$

To show this we note that

$$\log P^0(Y = 1 \mid A,W) = m(A,W \mid \beta^0 + \epsilon) + r^0(W) + \epsilon r(W)$$

so that

$$\frac{d}{d\epsilon} \log P_r^0(\epsilon)(Y = 1 \mid A,W)\Big|_{\epsilon=0} = \frac{d}{d\beta^0} m(A,W \mid \beta^0) + r(W).$$

For $Y = 0$, we have

$$\log P_r^0(\epsilon)(Y = 0 \mid A,W) = \log(1 - \exp(m(A,W \mid \beta^0 + \epsilon) + r^0(W) + \epsilon r(W)))$$

so that its derivative w.r.t. ϵ at $\epsilon = 0$ is given by

$$\frac{-1}{1 - Q^0(A,W)} Q^0(A,W)(\frac{d}{d\beta^0} m(A,W \mid \beta^0) + r(W)).$$

Thus, the score $S(r)$ is given by:

$$\begin{aligned} S(r) &= Y\left\{\tfrac{d}{d\beta^0} m(A,W \mid \beta^0) + r(W)\right\} - (1-Y)\tfrac{Q^0(A,W)}{1-Q^0(A,W)}\left\{\tfrac{d}{d\beta^0} m(A,W \mid \beta^0) + r(W)\right\} \\ &= \tfrac{Y}{1-Q^0(A,W)}\left\{\tfrac{d}{d\beta^0} m(A,W \mid \beta^0) + r(W)\right\} - \tfrac{Q^0(A,W)}{1-Q^0(A,W)}\left\{\tfrac{d}{d\beta^0} m(A,W \mid \beta^0) + r(W)\right\} \\ &= \tfrac{1}{1-Q^0(A,W)}\left\{\tfrac{d}{d\beta^0} m(A,W \mid \beta^0) + r(W)\right\}(Y - Q^0(A,W)). \end{aligned}$$

Following the strategy, we now have to select a $r(W)$ so that this score $S(r)$ is an element of (4 in which case it has to be equal (up till a normalizing constant matrix) to the efficient influence curve.

Thus we need that $m^*(A,W \mid \beta^0)h(A \mid W) = \frac{1}{1-Q^0(A,W)}\{\frac{d}{d\beta^0} m(A,W \mid \beta^0) + r(W)\}$ for some $h(A \mid W)$ with $E(h \mid W) = 0$. Thus, we need that $\frac{1}{m^*(1-Q^0(A,W))}(\frac{d}{d\beta^0} m(A,W \mid \beta^0) + r)$ has conditional mean zero, given W. (We remind the reader that $m^*(A,W \mid \beta^0) = Q^0(0,W)/Q^0(A,W)$.) Thus, it follows that

$$\begin{aligned} r(Q^0,g)(W) &= -\frac{E_g\left(\frac{1}{m^*(A,W|\beta^0)(1-Q^0(A,W))}\frac{d}{d\beta^0} m(A,W \mid \beta^0) \mid W\right)}{E_g\left(\frac{1}{m^*(A,W|\beta^0)(1-Q^0(A,W))} \mid W\right)} \\ &= -\frac{E_g\left(\frac{Q^0(A,W)}{1-Q^0(A,W)}\frac{d}{d\beta^0} m(A,W \mid \beta^0) \mid W\right)}{E_g\left(\frac{Q^0(A,W)}{(1-Q^0)(A,W)} \mid W\right)} \end{aligned}$$

This function r corresponds with

$$h_{opt}(Q^0, g) = \frac{1}{m^*(1-Q)} \left\{ \frac{d}{d\beta} m + r(Q^0, g) \right\}.$$

The efficient influence curve at $P_{Q^0,g}$ can thus be represented as

$$\begin{aligned} &D^*(Q^0, g)(W, A, Y) = m^*(W, A) h_{opt}(Q^0, g)(W, A)(Y - Q^0(W, A)) \\ &= h_{opt}(Q^0, g)(W, A)(Y m^*(W, A) - Q(0, W)) \\ &\equiv D_{h_{opt}(Q^0,g), q_{opt}(Q^0)}(g, \beta)(W, A, Y), \end{aligned}$$

where $q_{opt}(Q) = Q(0, W)$, $D_{h,q}(g, \beta) = (h(A, W) - E_g(h(A, W) \mid W))(Y m^*(A, W \mid \beta) - q(W))$ is a class of unbiased estimating functions indexed by choices h, q representing a subset of the orthogonal complement of the nuisance tangent space in the model with g known. We have the double robustness

$$E_0 D_{h,q}(g, \beta_0) = 0 \text{ if } g = g_0 \text{ or } q = E(Y|A = 0, W).$$

The iterative targeted-MLE: This defines the wished ϵ-extension $Q^0(\epsilon)$ of an initial fit $Q^0(Y \mid A, W)$. For example, if $m(W, A|\beta) = \beta A$, then this ϵ-fluctuation corresponds with adding $\epsilon C(A, W)$ to the initial fit $\log Q^0(A, W) = \beta^0 A + r^0(W)$, where the covariate

$$C(A, W) = A - \frac{E_g\left(\frac{Q^0(A,W)}{1-Q^0(A,W)} A \mid W\right)}{E_g\left(\frac{Q^0(A,W)}{(1-Q^0)(A,W)} \mid W\right)}.$$

Let ϵ_n^0 be the MLE over ϵ for $Q^0(\epsilon)$. Let $Q_n^1 = Q_n^0(\epsilon_n^0)$ be the updated estimate of $E(Y \mid A, W)$, which corresponds with an updated β_n^1 and $Q_n^1(0, W)$. We iterate this updating process till the corresponding sequence β_n^k is such that $\beta_n^k - \beta_n^{k-1}$ does not significantly change anymore. We denote the selected final update with $Q_n = Q_n^{k*}$ for some k^*, and $\beta_n = \beta_n^{k*}$, respectively, and we refer to this estimate β_n as the (iterative) targeted MLE of β_0.

Statistical Inference: Let $h_n = h_{opt}(Q_n^{k^*}, g_n)$ and $q_n = q_{opt}(Q_n)$ be the with Q_n and g_n corresponding indices. We have that up till a negligible term

$$0 = \sum_i D_{h_n, q_n}(\beta_n, g_n)(O_i).$$

That is, β_n can be viewed as a solution of the double robust estimating function for an index h_n, q_n which correspond with the final MLE update. Therefore, statistical inference for β_n can be based on the influence curve for this estimating equation as in van der Laan, Robins (2003), under the assumption that g_n

is correctly specified. If one wishes to only rely on the double robustness of β_n w.r.t. to misspecification of g_n and Q_n, then we recommend the bootstrap for statistical inference.

The ϵ-extension for discrete outcomes using Poisson fluctuations: We now use a fluctuation of the Poisson regression model. So we have, using short-hand notation

$$\begin{aligned} \log Q &= m_\beta + g \\ \log Q(\epsilon) &= m_{\beta+\epsilon} + g + \epsilon r \\ \log P(\epsilon)(Y \mid A, W) &= \frac{Q(\epsilon)^Y}{Y!} \exp(-Q(\epsilon)) \end{aligned}$$

We need that the score of this Poisson-distribution fluctuation at $\epsilon = 0$ is an element of the orthogonal complement of the nuisance tangent space, so that it equals the efficient score. We have

$$\frac{d}{d\epsilon} \log P(\epsilon)|_{\epsilon=0} = \{\frac{d}{d\beta} m_\beta + r\}(Y - Q).$$

Since the orthogonal complement of the nuisance tangent space consists of functions $m^* h(Y - Q)$ indexed by functions $h(A, W)$ with conditional mean zero, given W, we need $m^*(A, W \mid \beta^0) h(A \mid W) = \{\frac{d}{d\beta^0} m(A, W \mid \beta^0) + r(W)\}$ for some $h(A \mid W)$ with $E(h \mid W) = 0$. It follows that

$$r(Q^0, g)(W) = -\frac{E_g\left(\frac{1}{m^*(A,W|\beta^0)} \frac{d}{d\beta^0} m(A, W \mid \beta^0) \mid W\right)}{E_g\left(\frac{1}{m^*(A,W|\beta^0)} \mid W\right)}$$

This function r corresponds with

$$h_{opt}(Q^0, g) = \frac{1}{m^*} \left\{ \frac{d}{d\beta} m + r(Q^0, g) \right\}.$$

The efficient influence curve at $P_{Q^0,g}$ can thus be represented as

$$\begin{aligned} &D^*(Q^0, g)(W, A, Y) = m^*(W, A) h_{opt}(Q^0, g)(W, A)(Y - Q^0(W, A)) \\ &= h_{opt}(Q^0, g)(W, A)(Y m^*(W, A) - Q(0, W)) \\ &\equiv D_{h_{opt}(Q^0,g), q_{opt}(Q^0)}(g, \beta)(W, A, Y), \end{aligned}$$

where $q_{opt}(Q)(W) = Q(0, W)$, $D_{h,q}(g, \beta) = (h(A, W) - E_g(h(A, W) \mid W))(Y m^*(A, W \mid \beta) - q(W))$ is a class of unbiased estimating functions indexed by choices h, q

representing a subset of the orthogonal complement of the nuisance tangent space in the model with g known. We have the double robustness

$$E_0 D_{h,q}(g, \beta_0) = 0 \text{ if } g = g_0 \text{ or } q = E(Y|A = 0, W).$$

Thus, using this fluctuation function corresponds with adding a clever covariate (for model $m_\beta = \beta A$) given by

$$C(A, W) = A - \frac{E_g(A/m * |W)}{E_g(1/m * |W)},$$

where $m * (A, W) = E(Y|A = 0, W)/E(Y|A, W) = \exp(-m_\beta)$ is identified by the model m_β for $logE(Y|A, W)/E(Y|A = 0, W)$.

10 The targeted MLE for variable importance and causal effect, while allowing for missing outcome, missing treatment, or missing effect mofifier.

Suppose that we observe on each experimental unit $O = (W^*, \Delta, \Delta(Y, A, V))$, where we assume the missing at random assumption $\Pi(W^*) \equiv P(\Delta = 1 \mid (W, A, Y)) = P(\Delta = 1 \mid W^*)$. The likelihood contains the following factors:$P(Y \mid A, W)$, $P(A \mid W)$, $P(\Delta \mid W)$, $P(W)$:

$$P(O) = \{P(W)P(A \mid W)P(Y \mid A, W, \Delta = 1)\}^{\Delta} \{P(W^*)\}^{1-\Delta}.$$

In each of the above applications we represented the efficient influence curve of the parameter of interest based on (W, A, Y) as $D = D_1 + D_2$, where for a particular function h^* $D_1(W, A, Y) = h^*(A, W)(Y - Q_0(A, W))$ (with $Q_0(A, W) = E_0(Y \mid A, W)$) is the component of the efficient influence curve which corresponds with a score for $P(Y \mid A, W)$. This fact implied an ϵ-extension of the form $\epsilon h^*(A, W)$.

The efficient influence curve for this more general missing data structure can be represented as

$$\begin{aligned} D^* &= \frac{\Delta}{P(\Delta = 1 \mid W^*)}(D_1 + D_2) - E(D_1 + D_2 \mid \Delta = 1, W)\frac{\Delta}{P(\Delta = 1 \mid W^*)} \\ &\quad + E(D_1 + D_2 \mid \Delta = 1, W). \end{aligned}$$

The component of D^* corresponding with the likelihood factor $P(Y \mid \Delta = 1, A, W)$ is therefore given by:

$$D_1^* = \frac{\Delta}{\Pi(W^*)} D_1 = \frac{\Delta}{\Pi(W^*)} h^*(A, W)(Y - E(Y \mid A, W, \Delta = 1)).$$

Thus one needs to arrange an ϵ extension $Q^0(\epsilon)(Y \mid A, W, \Delta = 1)$ of an initial fit $Q^0(Y \mid A, W, \Delta = 1)$ which has a score at $\epsilon = 0$ given by D_1^*. As a consequence, we can use the same epsilon-extensions as proposed above as basis, BUT now restricted to the observations with $\Delta = 1$ and we should multiply the ϵ-covariate with $1/\Pi(W)$.

Regarding statistical inference, we should use that the targeted MLE is a solution of $0 = \sum_i D^*(\beta_n, Q_n, g_n, \Pi_n)(O_i)$ with D^* defined above so that it can be analyzed as the double robust estimator. Note also that for fitting $g(A \mid W)$ one should note $P(A = a \mid W) = P(A = a \mid W, \Delta = 1)$ so that also the treatment mechanism (just like $E(Y \mid A, W) = E(Y \mid A, W, \Delta = 1)$) can be fitted by just restricting to the observations with $\Delta = 1$.

This generalization of the targeted MLE to missing data also applies to the examples below.

11 Targeted MLE for a marginal structural logistic regression model for survival outcome.

For each treatment choice $a \in \mathcal{A}$, let $T(a)$ be a treatment specific counterfactual survival time, and let the full data on each experimental unit be given by $(W, (T(a) : a \in \mathcal{A}))$. Suppose we observe $O = (W, A, T = T(A))$. Suppose that the survival times are discrete on time points indexed by $j = 0, 1, \ldots$. Consider the following class of causal working models for the treatment specific hazard:

$$P(T(a) = t \mid T(a) \geq t, V) = m(a, t, V \mid \beta_0),$$

for a given working model $m(a, t, V \mid \beta)$ indexed by parameter vector β. Let $dN(t) = I(T = t)$ and $dN_a(t) = I(T(a) = t)$. The typical working model will be a logistic regression model:

$$m(a, t, V \mid \beta) = \frac{1}{1 + \exp(-m_0(a, t, V \mid \beta))},$$

where m_0 is a specified function linear in summary measures of (a, t, V). We also assume the randomization assumption: A is independent of X, given W.

The class of so called IPTW-estimating functions for β_0 are given by:

$$D_{IPTW,h} = \frac{1}{g_0(A \mid X)} \sum_t h(A,t,V) \frac{d}{d\beta_0} m(A,t,V \mid \beta_0)(dN(t) - I(T \geq t) m(A,t,V \mid \beta_0)).$$

By projecting the $D_{IPTW,h}$ on the tangent space of the relevant (i.e., ignoring the treatment mechanism) factor of the likelihood of O, given by

$$P(W) \prod_t P(dN(t) \mid \bar{N}(t-1), A, W),$$

we obtain the efficient influence curve of $\beta_0 = \beta_{0h}$ defined non-parametrically as the solution of $P_0 D_{IPTW,h}(\beta, g_0) = 0$.

Let $q_0(t \mid A, W) = P_0(T = t \mid A, W)$, and $\bar{Q}_0(t \mid A, W) = P_0(T \geq t \mid A, W)$. We have the following representation of this efficient influence curve

$$\begin{aligned} &D_h^*(\beta_0, Q_0^r, g^r) = \\ &= \textstyle\sum_{t=0}^T h^*(Q_0^r, g^r)(t, \bar{A}_1(t-1), W)(dY(t) - E(dY(t) \mid \bar{Y}(t-1), \bar{A}_1(t-1), W)) \\ &+ E(D_{IPTW} \mid W) \\ &\equiv D_1(Q_0^r, g_0^r)(W, A, Y) + D_2(Q_0^r)(W), \end{aligned}$$

where

$$\begin{aligned} h^* = E_{Q_0^r, g_0^r}(D_{IPTW} \mid dY(t) = 1, \bar{Y}(t-1), A_1, W) - \\ E_{Q_0^r, g_0^r}(D_{IPTW} \mid dY(t) = 0, \bar{Y}(t-1), A_1, W), \end{aligned}$$

where D_2 represents a score of the marginal distribution of W and the first term D_1 represents a sum over t of scores of $P(dY(t) \mid \bar{Y}(t-1), A_1, W)$. In the special case that $W = V$, we have that $D_h^* = D_{IPTW,h}$ since $D_{IPTW,h}$ is already an element of the tangent space.

The second term defines $\beta(Q)$ as a function of Q through the following least squares solution (check):

$$\beta(Q) = \arg\min_\beta E_Q \sum_a \sum_t h(t, \bar{a}_1(t-1), V) \left\{ q(t, a_1, V) - \bar{Q}(t, a_1, V) \lambda_\beta(t, a_1, V) \right\}^2,$$

and for Q_{1n} being the empirical distribution of $W_1, \ldots, W_n$ this gives us:

$$\beta(Q_{1n}, Q_{2n}) = \arg\min_\beta \sum_i \sum_a \sum_t h(a_1, t, V_i) \left\{ q_n(t, a_1, V_i) - \bar{Q}_n(t, a_1, V_i) \lambda_\beta(t, a_1, V_i) \right\}^2.$$

In other words, the choice of h defines β_0 as a weighted projection of the true hazard $q_0/\bar{Q}_0$ on the working model $\{\lambda_\beta(a_1, t, V) : \beta\}$.

Consider an initial fit of $\lambda^0(t \mid A, W)$ of $E(dN(t) \mid \bar{N}(t-1) = 0, A, W)$ based on a logistic regression model, and represent it as follows:

$$\lambda^0(t \mid A, W) = \frac{1}{1 + \exp(-m^0(t, A, W))}.$$

Consider the following ϵ-extension:

$$\lambda^0(\epsilon)(t \mid A, W) = \frac{1}{1 + \exp(-m^0(t, A, W) - \epsilon h^*(t, A, W))}.$$

The score of $\lambda^0(\epsilon)$ at $\epsilon = 0$ equals the wished component $D_2(\beta(Q^0), Q^0, g_0)$ of the efficient influence curve. Thus, assuming an initial fit Q^0 for which Q_1^0 is the empirical distribution of $W_1, \ldots, W_n$, it follows that the with $\lambda^0(\epsilon)$ corresponding $Q^0(\epsilon)$ (and no update of the already nonparametric MLE Q_1^0) has score at $\epsilon = 0$ equal to the efficient influence curve $D^*(\beta^0, Q^0, g_0)$.

The iterative targeted-MLE: This defines the wished ϵ-extension $Q^0(\epsilon)$ of an initial fit Q^0. Let ϵ_n^0 be the MLE over ϵ for $Q^0(\epsilon)$. Let $Q_n^1 = Q_n^0(\epsilon_n^0)$ be the updated estimate which corresponds with an updated $\beta_n^1 = \beta_h(Q_n^1)$. We iterate this updating process till the corresponding sequence β_n^k is such that $\beta_n^k - \beta_n^{k-1}$ does not significantly change anymore. In the case that $h = h_1/(m(1-m))$ is chosen so that r^* does not depend on β, then it follows that this iterative targeted MLE converges in one step.

We denote the selected final update with $Q_n = Q_n^{k*}$ for some k^*, and $\beta_n = \beta_n^{k*}$, respectively, and we refer to this estimate β_n as the (iterative) targeted MLE of β_0.

Statistical Inference: We have that up till a negligible term

$$0 = \sum_i D_{h_n}^*(\beta_n, Q_n, g_n)(O_i).$$

That is, β_n can be viewed as a solution of the double robust estimating function for an index h_n. Therefore, statistical inference for β_n can be based on the influence curve for this estimating equation as in van der Laan, Robins (2003), under the assumption that g_n is correctly specified. If one wishes to only rely on the double robustness of β_n w.r.t. to misspecification of g_n and Q_n, then we recommend the bootstrap for statistical inference.

12 Template for proving Asymptotic Linearity of Targeted MLE.

In this section we show how one can establish asymptotic linearity of targeted MLE for the target parameter of interest without having to use that the gradi-

ent or canonical gradient of the path-wise derivative can be represented as an estimating function for the parameter of interest, as in van der Laan, Robins (2002).

Let P_n^* be a targeted MLE so that $P_n D(P_n^*) = 0$, where $D(P)$ is a gradient of the path-wise derivative of the target parameter $\Psi : \mathcal{M} \to \mathbb{R}^d$ at P. We have

$$\begin{aligned} \Psi(P_n^*) &= \Psi(P_n^*) + P_n D(P_n^*) \\ &= \Psi(P_n^*) + (P_n - P_0)D(P_n^*) + P_0 D(P_n^*). \end{aligned}$$

In general, one will need to establish that

$$P_0 D(P_n^*) = \psi_0 - \Psi(P_n^*) + (P_n - P_0)D_1(P_0) + o_P(1/\sqrt{n}) \tag{5}$$

for some mean zero function $D_1(P_0)(O)$. A special case is that $D_1 = 0$ which, by path-wise differentiability, one expects to hold if $D(P_n^*)$ is a consistent estimator of $D(P_0)$. For example, if the model $\mathcal{M}$ is convex, Ψ is linear, and dP_0/dP_n^* exists, then $P_0 D(P_n^*) = \psi_0 - \Psi(P_n^*)$ exact (i.e., no remainder) (van der Laan, 1996). Given this assumption (6, one obtains

$$\Psi(P_n^*) - \psi_0 = (P_n - P_0)\left\{D(P_n^*) + D_1(P_0)\right\} + o_P(1/\sqrt{n}).$$

Under empirical process conditions on $D(P_n^*)$, and that $D(P_n^*)$ converges to some $D(P_0^*)$ for a $P_0^* \in \mathcal{M}$ (not necessarily equal to P_0), one now obtains the wished asymptotic linearity

$$\Psi(P_n^*) - \psi_0 = (P_n - P_0)\{D(P_0^*) + D_1(P_0)\} + o_P(1/\sqrt{n}).$$

Theorem 4 *Consider a sample of n i.i.d. observations $O_1, \ldots, O_n \sim P_0$, where P_0 is known to be an element of model $\mathcal{M}$. Let $\Psi : \mathcal{M} \to \mathbb{R}^d$ be a Euclidean target parameter of interest. Let P_n^* be an estimator of P_0 satisfying $P_n D(P_n^*) = 1/n \sum_i D(P_n^*)(O_i) = 0$, where $D(P)$ is a gradient of the path-wise derivative at P of the target parameter $\Psi : \mathcal{M} \to \mathbb{R}^d$.*

- *Assume*

$$P_0 D(P_n^*) = \psi_0 - \Psi(P_n^*) + (P_n - P_0)D_1(P_0) + o_P(1/\sqrt{n}) \tag{6}$$

for some mean zero function $D_1(P_0)$ of O. Under this assumption one obtains

$$\Psi(P_n^*) - \psi_0 = (P_n - P_0)\left\{D(P_n^*) + D_1(P_0)\right\} + o_P(1/\sqrt{n}).$$

- *In addition, assume $D(P_n^*)$ falls in a P_0-Donsker class with probability tending to 1. Then, $\Psi(P_n^*) - \psi_0 = O_P(1/\sqrt{n})$.*
- *In addition, assume $P_0\{D(P_n^*) - D(P_0^*)\}^2 \to 0$ in probability as $n \to \infty$ for some $D(P_0^*)$ in the P_0-Donsker class.*

Then,

$$\Psi(P_n^*) - \psi_0 = (P_n - P_0)\{D(P_0^*) + D_1(P_0)\} + o_P(1/\sqrt{n}).$$

In particular, if $D_1(P_0) = 0$ and $D(P_0^*) = D^*(P_0)$ equals the canonical gradient at P_0, then $\Psi(P_n^*)$ is asymptotically efficient.

Consider now CAR-censored data models so that $D(P) = D(Q(P), g(P))$, $\Psi(P)$ depends on P through $Q(P)$ only, and the density factorizes as $p = Q(p)g(p)$. Let $p_n^* = Q_n^* g_n^*$. We consider the case that g_n^* is assumed to be consistent for g_0. Regarding verification of (6), we can exploit some structure. That is, we proceed as follows:

$$\begin{aligned} P_0 D(Q_n^*, g_n^*) &= P_0 D(Q_n^*, g_0) - \{P_0 D(Q_n^*, g_n^*) - D(Q_n^*, g_0)\} \\ &= P_0 D(Q_n^*, g_0) + P_0\{D(Q_0^*, g_n^*) - D(Q_0^*, g_0)\} \\ &\quad -R_{1n} \end{aligned}$$

where

$$R_{1n} = P_0\{D(Q_n^*, g_n^*) - D(Q_n^*, g_0)\} - P_0\{D(Q_0^*, g_0) - D(Q_0^*, g_0)\}.$$

Here R_{1n} is a second order term and therefore it is natural to make it an assumption that $R_{1n} = o_P(1/\sqrt{n})$. Secondly, we define

$$\Phi(g_n^*) \equiv P_0 D(Q_0^*, g_n^*),$$

so that the term $P_0\{D(Q_0^*, g_n^*) - D(Q_0^*, g_0)\}$ equals $\Phi(g_n^*) - \Phi(g_0)$. We now assume that $\Phi(g_n^*)$ is an efficient estimator of the parameter $\Phi(g_0)$ in model $\mathcal{M}(\mathcal{G}) = \{p_{Q,g} = Qg : g \in \mathcal{G}\}$, where we denote the tangent space generated by model $\mathcal{G}$ for g_0 at $P_0 = Q_0 g_0$ with $T_g(P_0)$. It remains to consider $P_0 D(Q_n^*, g_0)$. By the general representation Theorem 1.3 in van der Laan, Robins (2002), it follows that

$$P_0 D(Q_n^*, g_0) = P_{Q_0} D^F(Q_n^*),$$

where $D^F(Q)$ is a gradient in the full data model $\mathcal{Q}$ for the parameter $Q \to \Psi(Q)$, and P_{Q_0} denotes the full data distribution. Again, by path-wise differentiability of Ψ in the full data model, if $D^F(Q_n^*)$ consistently estimates

$D^F(Q_0)$, then one expects $P_{Q_0}D^F(Q_n^*) = \psi_0 - \Psi(Q_n^*) + o_P(1/\sqrt{n})$. In general, we note that, if Q_n^* converges to some possibly misspecified Q^* for which $\Psi(Q^*) = \Psi(Q_0)$ and $P_0D^F(Q^*) = 0$, we have

$$P_{Q_0}D^F(Q_n^*) = P_{Q^*}D^F(Q_n^*) + P_{Q_0-Q^*}\{D^F(Q_n^*) - D^F(Q^*)\}.$$

By pathwise differentiabiliy, and the convergence of Q_n^* to Q^* the first order Tailor expansion suggests

$$P_{Q^*}D^F(Q_n^*) = \psi_0 - \Psi(Q_n^*) + o_P(1/\sqrt{n}).$$

A separate study of the other term (which can be represented as $\Phi(Q_n^*) - \Phi(Q^*)$ for some Φ) will result in an asymptotic linearity result:

$$P_{Q_0-Q^*}\{D^F(Q_n^*) - D^F(Q^*)\} = (P_n - P_0)D_1(P_0) + o_P(1/\sqrt{n}).$$

To stay general, we assume the expansion (6):

$$P_0D(Q_n^*, g_0) = \psi_0 - \Psi(Q_n^*) + \frac{1}{n}\sum_{i=1}^{n} D_1(P_0) + o_P(1/\sqrt{n}),$$

for some $D_1(P_0)$. By Theorem 2.3 in van der Laan, Robins (2002) the influence curve of $\Phi(g_n^*)$ equals $-\Pi(D(Q_0^*, g_0) + D_1(P_0) \mid T_g(P_0)^\perp)$. This proves the following theorem which provides a template for establishing asymptotic linearity of the targeted MLE in CAR censored data models.

Theorem 5 *Let $O_1, \ldots, O_n \sim P_0$ be n i.i.d. copies of $O = \Phi(C, X)$ for some many to one mapping Φ of censoring variable C and full data structure X. Assume that the conditional distribution G_0 of C, given X, satisfies CAR so that $p_0 = Q_0g_0$ w.r.t to appropriate dominating measure, g_0 is a density of G_0 and Q_0 a function of distribution of full data X. Let $\mathcal{M} = \{p_{Qg} = Qg : Q \in \mathcal{Q}, g \in \mathcal{G}\}$, where $\mathcal{G}$ is a subset of all CAR distributions. Let $\Psi : \mathcal{Q} \to \mathbb{R}^d$ be the Euclidean target parameter of interest. Let $D(P) = D(Q(P), g(P))$ be a gradient of Ψ at $P \in \mathcal{M}$. Consider an estimator P_n^* with density $p_n^* = Q_n^*g_n^*$ satisfying $P_nD(Q_n^*, g_n^*) = 0$.*

- *Define*

$$R_{1n} \equiv P_0\{D(Q_n^*, g_n^*) - D(Q_n^*, g_0)\} - P_0\{D(Q_0^*, g_0) - D(Q_0^*, g_0)\}.$$

Assume $R_{1n} = o_P(1/\sqrt{n})$.

- *Define*
$$\Phi(g_n^*) \equiv P_0 D(Q_0^*, g_n^*),$$
where P_0 and Q_0^* are treated as given. Assume that $\Phi(g_n^*)$ is an efficient estimator of the parameter $\Phi(g_0)$ in model $\mathcal{M}(\mathcal{G}) = \{p_{Q,g} = Qg : Q \in \mathcal{Q}, g \in \mathcal{G}\}$, and let $T_g(P_0)$ denote the tangent space generated by model $\mathcal{G}$ for g_0 at $P_0 = Q_0 g_0$.

- *Assume the expansion (6):*
$$P_0 D(Q_n^*, g_0) = \psi_0 - \Psi(Q_n^*) + \frac{1}{n}\sum_{i=1}^{n} D_1(P_0) + o_P(1/\sqrt{n}),$$
for some $D_1(P_0)$.

- *Assume $D(Q_n^*, g_n^*)$ falls in a P_0-Donsker class. Then, $\Psi(P_n^*) - \psi_0 = O_P(1/\sqrt{n})$.*

- *In addition, assume $P_0\{D(Q_n^*, g_n^*) - D(Q_0^*, g_0)\}^2 \to 0$ in probability as $n \to \infty$ for some Q_0^* and $D(Q_0^*, g_0)$ in the P_0-Donsker class.*

Then,
$$\Psi(P_n^*) - \psi_0 = (P_n - P_0) IC(P_0) + o_P(1/\sqrt{n}),$$
where
$$IC(P_0) \equiv \Pi\left(D(Q_0^*, g_0) + D_1(P_0) \mid T_g(P_0)^{\perp}\right),$$
Π is the projection operator in $L_0^2(P_0)$ endowed with inner product $\langle f, g\rangle_{P_0} = E_{P_0} fg$ onto the orthogonal complement of $T_g(P_0)$. If $D_1(P_0) = 0$ and $D(Q_0^*, g_0) = D^*(Q_0, g_0)$ where D^* is the canonical gradient, then $\Psi(P_n^*)$ is asymptotically efficient.

13 Targeted MLE for causal effect of treatment on survival outcome allowing for right-censoring.

In this section we illustrate how one can also apply the targeted MLE to deal with complex longitudinal data structures: as pointed out in van der Laan, Rubin (2006), the targeted MLE has an analogue of the double robust estimators presented in van der Laan, Robins (2004) for each model, parameter and censored data structure. This example is complex enough so that it becomes

clear how one can immediately generalize the presented t-MLE to longitudinal time dependent treatments and time-dependent covariate processes.

Suppose that the full data of interest consists of baseline covariates W, and a set of treatment specific survival times $T_{a(0)}$ with support $\{0, 1, \ldots, \tau + 1\}$ indexed by a set of possible single time point treatments $a(0)$ assigned at baseline. Let $a = (a(0), a(1), \ldots, a(\tau))$ with $a(t) = I(c = t)$, $t = 1, 2, \ldots, \tau$ for some set censoring time c: thus, $a(t)$ has only a single 1 at most, and after this 1 it stays zero. If $a(1) = \ldots = a(\tau) = 0$, then we will also refer to this as $c = \infty$. Let $L(0) = W$, $L_a(t) = (I(T_{a(0)} \leq \min(t, c)), t = 1, \ldots$, where $L_a(t)$ can also be represented as $I(\tilde{T}_a \equiv \min(T_{a(0)}, c) \leq t)$. The full data is $X = (L_a : a \in \mathcal{A})$. The observed data on each experimental unit is $O_i = (A_i, L_{A_i})$, where A_i identifies the assigned treatment $A_i(0)$ and the right-censoring time C_i, where $C_i \equiv \infty$ if $T_i \leq C_i$. Equivalently, $O_i = (W_i, A_i, \tilde{T}_i = \tilde{T}_{A_i})$. Let

$$\psi_0(t) = P(T_{a(0)} > t) - P(T_0 > t) = P(\tilde{T}_{a(0)0} > t) - P(\tilde{T}_{00} > t),$$

where $\tilde{T}_{a(0)0} = T_{a(0)}$ is the follow up time if censoring is set at $c = \infty$, which thus equals the treatment specific survival time $T_{a(0)}$.

A CAR design is a conditional distribution of A, given X, satisfying

$$\begin{aligned} g(a \mid X) &= \prod_{t=0}^{\tau} g(a(t) \mid \bar{A}(t-1) = \bar{a}(t-1), X) \\ &\stackrel{CAR}{=} \prod_{t=0}^{\min(c(a), T_a - 1)} g(a(t) \mid \bar{A}(t-1) = \bar{a}(t-1), \bar{L}_A(t)) \prod_{t=T_a}^{\tau} I(a(t) = 0), \end{aligned}$$

where the first factor at $t = 0$ denotes a treatment mechanism, and the other factors represent the censoring mechanism, where censoring is set at ∞ after the failure time T_a.

The density of $O_1, \ldots, O_n$ is given by:

$$\begin{aligned} p_{Q_0,g}(O_1, \ldots, O_n) &= \prod_{i=1}^{n} Q_0(A_i, L_i) g(A_i \mid X_i) \\ &= \prod_{i=1}^{n} \prod_{t=0}^{\tau+1} Q_{0t}(L_i(t) \mid \bar{L}_i(t-1), \bar{A}_i(t-1)) \prod_{i=1}^{n} g(A_i \mid X_i) \\ &= \prod_{i=1}^{n} \prod_{t=0}^{\tilde{T}_i = \min(T_{iA_i}, C_i)} Q_{0t}(L_i(t) \mid \bar{L}_i(t-1), \bar{A}_i(t-1)) \\ &\quad \prod_{i=1}^{n} \prod_{t=0}^{\tilde{T}_i} g_t(A_i(t) \mid \bar{A}_i(t-1), \bar{L}_i(t)), \end{aligned}$$

where $Q_{0t}(l(t) \mid \bar{l}(t-1), \bar{a}(t-1))$ denotes the conditional probability distribution of $L_a(t)$ at $l(t)$ given $\bar{L}_a(t-1) = \bar{l}(t-1)$, so that $Q_0(a,l) = P(L_a = l)$. Note that Q_{0t} models the hazard of survival time $T_i = T_{A_i(0)}$, given T_i has not happened yet, baseline covariates W_i, and treatment $A_i(0)$.

We could model Q_{0t} with a logistic regression model:

$$Q_{t\theta}(dL(t) = 1 \mid \bar{L}(t-1), L(0), \bar{A}(t-1), A(0)) = I(\tilde{T} \geq t) \frac{1}{1 + \exp(-\theta(t) f(L(0), A(0)))},$$

where $f(L(0), A(0))$ is a vector valued summary measure of $(L(0), A(0))$. Note that this logistic function at $\bar{A}(t-1) = \bar{a}(t-1)$ is actually modelling

$$P(dL(t) = 1 \mid \bar{L}(t-1), \bar{A}(t-1) = \bar{a}(t-1), \tilde{T} \geq t) = P(T_{a(0)} = t \mid T_{a(0)} \geq t, L(0)),$$

which is the hazard of the treatment specific survival time $T_{a(0)}$ indexed by baseline treatment $a(0)$, conditional on $L(0)$. Let $\mathcal{Q}^w = \{Q_\theta : \theta\} = \{(Q_{t\theta(t)} : t) : \theta\}$ be this model for Q_0. The maximum likelihood estimator of $\theta = (\theta(t) : t)$ can be computed with standard logistic regression software. If the survival time is continuous (e.g.,, the time scale is chosen fine enough so that no ties occur at the same time), then one could also model Q_{0t} with a multiplicative intensity model:

$$P(dL(t) = 1 \mid \bar{L}(t-), \bar{A}(t-)) = I(\tilde{T} \geq t) \lambda_0(t) \exp(\theta f_t(L(0), A(0))),$$

where $f_t(L(0), A(0))$ is a time dependent covariate defined as a function of $t, L(0), A(0)$. In particular, one could assume a Cox-proportional hazards model for T conditional on $L(0), A(0)$. Again, the maximum likelihood estimator for the multiplicative intensity model can be fitted with standard software (e.g., Coxph() in R).

We now wish to consider the targeted maximum likelihood estimator of ψ_0 based on this maximum likelihood estimator for this working model $\mathcal{Q}^w$.

An inverse probability of censoring weighted estimating function for $\psi_0(t_0)$ is given by:

$$\begin{aligned} D_{IPCW}(Q_0, g)(O) &= I(T > t_0) I(C \geq t_0) \left\{ \frac{I(A(0) = a(0))}{g(a(0)\bar{0}(t_0) \mid X)} - \frac{I(A(0) = 0)}{g(0\bar{0}(t_0) \mid X)} \right\} \\ &\quad - \Psi(Q_0)(t_0), \end{aligned}$$

where

$$g(a(0)\bar{0}(t_0) \mid X) = g(a(0) \mid L(0)) \prod_{t=1}^{t_0} P(A(t) = 0 \mid A(0) = a(0), A(t-1) = 0, L(0))$$

is the conditional probability of having $a(0)$ assigned and being uncensored up till and including time t_0.

The efficient influence curve $D^*(Q_0, g)$ at a data generating distribution $P_{Q_0,g}$ can be represented as the projection of D_{IPCW} onto the tangent space of the Q_0-factor of the likelihood/density of the experimental unit's data structure O:

$$\begin{aligned} D^*(Q_0,g) &= \Pi(D_{IPCW}(Q_0,g) \mid T(Q_0)) \\ &= \sum_{t=0}^{\tau+1} E_{Q_0,g}(D_{IPCW}(Q_0,g) \mid \bar{L}(t), \bar{A}(t-1)) \\ &\quad - \sum_{t=0}^{\tau+1} E_{Q_0,g}(D_{IPCW}(Q_0,g) \mid \bar{L}(t-1), \bar{A}(t-1)) \\ &= \sum_{t=0}^{\tilde{T}} E_{Q_0,g}(D_{IPCW}(Q_0,g) \mid L(0), A(0), dL(t), \tilde{T} = \min(T,C) \geq t) \\ &\quad - \sum_{t=0}^{\tilde{T}} E_{Q_0,g}(D_{IPCW}(Q_0,g) \mid L(0), A(0), \tilde{T} = \min(T,C) \geq t), \end{aligned}$$

where we recall that $dL(t) = I(\tilde{T} = t, C \geq t)$ equals 1 if a failure $T = t$ occurs at time t (but $dL(t) = 0$ if $C = t$ but $T \neq t$). We define $h^*(Q_0, g)(t, L(0), A(0))$ as

$$\begin{aligned} &E_{Q_0,g}(D_{IPCW}(Q_0,g) \mid L(0), A(0), dL(t) = 1, \tilde{T} = \min(T,C) \geq t) \\ &-E_{Q_0,g}(D_{IPCW}(Q_0,g) \mid L(0), A(0), dL(t) = 0, \tilde{T} = \min(T,C) \geq t), \end{aligned}$$

and we note that

$$\begin{aligned} D^*(Q_0,g) &= \sum_{t=0}^{\tilde{T}} D_t^*(Q_0,g) \\ &= \sum_{t=0}^{\tilde{T}} h^*(Q_0,g)(t, L(0), A(0))(dL(t) - Q_{0t}(1 \mid \bar{L}(t-1), \bar{A}(t-1))). \end{aligned}$$

ϵ-fluctuation to define targeted MLE: Let $Q_{t\theta_n}(\epsilon)$ be a 1-dimensional extension parameter ϵ so that $Q_{t\theta_n}(0) = Q_{t\theta_n}$ and the score of ϵ at $\epsilon = 0$ for observation O_i equals $D_t^*(Q_{\theta_n}, g) = h^*(Q_{\theta_n}, g)(t, L_i(0), A_i(0))(dL_i(t) - Q_{t\theta_n}(1 \mid \bar{L}_i(t-1), \bar{A}_i(t-1)))$. This can be achieved by adding to the logistic regression model $Q_{t\theta_n}(1 \mid \bar{L}_i(t-1), \bar{A}_i(t-1))$ a covariate $h^*(Q_{\theta_n}, g)(t, L_i(0), A_i(0))$ with coefficient ϵ. Similarly, one can add this covariate to the multiplicative

intensity model. Let ϵ_n be the maximum likelihood estimator of ϵ for this one dimensional parametric model

$$\epsilon_n = \arg\max_{\epsilon} \prod_{i=1}^{n} \prod_{t} Q_{t\theta_n}(\epsilon)(L_i(t) \mid \bar{L}_i(t-1), \bar{A}_i(t-1)).$$

Computing ϵ_n corresponds with fitting a logistic regression model based on a pooled (across time) sample with a single regression coefficient ϵ and can thus be done with standard software. Let $Q_{\theta_n}(\epsilon_n) = (Q_{t\theta_n}(\epsilon_n) : t)$. The first step targeted maximum likelihood estimator of ψ_0 for the fixed design is now defined as $\psi_n = \Psi(Q_{\theta_n}(\epsilon_n))$. The k-th step targeted MLE is defined by iterating this process. One can also compute a one-step targeted MLE by defining ϵ_n as the solution of $0 = \sum_i D^*(Q_{\theta_n}(\epsilon), g) = 0$.

Application of our results in van der Laan, Rubin (2006) for the k-th step targeted MLE (k large enough) or this latter one step targeted MLE shows that, in the case that g_0 is known, ψ_n is consistent and asymptotically linear at P_{Q_0,g_0} with influence curve $D^*(Q^*, g_0)$, where Q^* is the limit of $Q_{\theta_n}(\epsilon_n)$. In particular, if $Q^* = Q_0$, i.e., if $\mathcal{Q}$ is correctly specified, then ψ_n is asymptotically efficient. In other words, the targeted MLE is locally efficient. If g_0 (including both treatment mechanism and censoring mechanism) is estimated with a maximum likelihood estimator according to a correctly specified model, then the above influence curve is generally conservative.

14 Targeted MLE for semi-parametric realistic MSM/V-adjusted additive variable importance

Consider the setting of our realistic marginal structural model for the causal effect of rules for point treatment. Recall the observed data structure $O = (W, A, Y)$ for the experimental unit. Consider parameter $\psi_0(a, V) = E(Y(d(a)) \mid V) - E(Y(d(0)) \mid V)$, and we refer to a model $\psi_0(a, V) = m(a, V \mid \beta_0)$ as a semi-parametric MSM for realistic point treatment interventions. The estimator of β_0 can also be used as an estimator of V-adjusted variable importance:

$$E(E(Y \mid A = d(a)(W), W) - E(Y \mid A = d(0)(W), W) \mid V).$$

Working model: We consider a working model $\{m(a, v \mid \beta) : \beta\}$ for $\psi_0(a, v)$, and define the target parameter as

$$\beta_0 = \arg\min_{\beta} E_{0V} \sum_{a \in \mathcal{A}_1} (m(a, V \mid \beta) - \psi_0(a, V))^2 h(a, V),$$

where h is a user supplied weight function. For simplicity, we assume here that $\mathcal{A}_1$ is discrete, but if $\mathcal{A}_1$ is a continuous set, then one can replace it by a discrete approximation in the above definition.

The first challenge is to determine the orthogonal complement of the nuisance tangent space $T_{nuis}^{\perp}(P)$ at a P in this working model, or a rich subset of this space. For this purpose we follow a strategy I proposed and used in Hubbard, van der Laan (2007, Population Intervention models) to derive a class of estimating functions for such semi-parametric causal models.

Firstly, we note that $E(Y(d(a)) \mid V) = m(a, V \mid \beta_0) + \theta_0(V)$, where $\theta_0(V) = E(Y(d(0)) \mid V) = E(E(Y \mid A = d(0)(W), W) \mid V)$. Let $\psi_0(a, V) = E(Y(d(a)) \mid V) - E(Y(d(0) \mid V)$.

We note that we also have

$$\beta_0 = \arg\min_{\beta} E_{0V} \sum_{a \in \mathcal{A}_1} (m(a, V \mid \beta) + \theta_0(V) - \theta_0(a, V))^2 h(a, V),$$

where $\theta_0(a, V) = E(Y(d(a)) \mid V)$. In the MSM-model with θ_0 known, the efficient influence curve for β_0 is given by

$$\begin{aligned} D^*(\beta_0, Q_0, g_0, \theta_0) &= \sum_{a \in \mathcal{A}_1} I(A = d(a)(W)) \frac{h(a, V) \frac{d}{d\beta_0} m(a, V \mid \beta_0)}{g_0(A \mid X)} (Y - Q_{02}(A, W)) \\ &+ \sum_{a \in \mathcal{A}_1} h(a, V) \frac{d}{d\beta_0} m(a, V \mid \beta_0)(Q_{02}(d(a)(W), W) - m(a, V \mid \beta_0) - \theta_0(V)) \\ &\equiv D_1^*(\beta_0, Q_0, g_0)(W, A, Y) + D_2^*(\beta_0, Q_0, \theta_0)(W), \end{aligned}$$

where we defined $Q_{02}(d(a)(W), W) = E_{Q_0}(Y \mid A = d(a)(W), W)$ and $Q_{02}(a, W) = E(Y \mid A = a, W)$, and we note that $\beta_0 = \beta(Q_0)$ is a parameter of $Q_0 = (Q_{01}, Q_{02})$.

Let θ_n be an estimator of θ_0. Let IC_{nu} be the influence curve of $-P_0 D^*(Q_0, g_0, \theta_n) - D^*(Q_0, g_0, \theta_0)$. Then under regularity conditions $-P_0(D^*(\beta_n) - D^*(\beta_0)) \approx (P_n - P_0)(D^* - IC_{nu})$. The estimating functions $D^* - IC_{nu}$ are the corrected estimating functions for the semiparametric model in which $\theta_0(V)$ is unspecified. Thus, we need to determine the influence curve IC_{nu} for a nonparametric estimator θ_n of θ_0.

We note that

$$-P_0(D(\theta_n) - D(\theta_0)) = P_0 \sum_{a \in \mathcal{A}_1} h(a, V) \frac{d}{d\beta_0} m(a, V \mid \beta_0)(\theta_n(V) - \theta_0(V)).$$

We will now work out a linearization for $\theta_n - \theta_0$. We have

$$E_n(Y \mid A = d(0)(w), W = w) = \frac{\sum_{i=1}^n Y_i I(A_i = d(0)(W_i), W_i = w)}{\sum_{i=1}^n I(A_i = d(0)(W_i), W_i = w)} = \frac{P_n Y I_w}{P_n I_w}.$$

Let $p_n(w \mid v)$ be an estimate of the conditional probability of W, given V. Then, $\theta_n(v) = \sum_w \frac{P_n Y I_w}{P_n I_w} p_n(w \mid v)$. Thus,

$$\begin{aligned}
&\theta_n(v) - \theta_0(v) = \\
&\textstyle\sum_w E_n(Y \mid A = d(0)(w), W = w) p_n(w \mid v) - E(Y \mid A = d(0)(w), W = w) p(w \mid v) \\
&= \textstyle\sum_{w^*} (E_n - E)(Y \mid A = d(0)(w^*, v), W = (w^*, v)) p(w^* \mid v) \\
&+ \textstyle\sum_{w^*} E(Y \mid A = d(0)(w^*, v), W = (w^*, v))(p_n - p)(w^* \mid v)
\end{aligned}$$

We have that

$$(E_n - E)(Y \mid A = d(0)(w), W = w) \approx \frac{P_n Y I_w}{P I_w} - \frac{\theta_0(w)}{P I_w} P_n I_w,$$

where we used the notation $\theta_0(w) = E(Y \mid A = d(0)(w), W = w)$. So the first term in the linearization of $\theta_n - \theta_0$ is given by:

$$\begin{aligned}
&\tfrac{1}{n} \textstyle\sum_i \sum_{w^*} \left\{ \frac{Y_i I(W_i = (w^*, v), A_i = d(0)(w^*, v))}{P I_w} - \frac{\theta_0(w^*, v)}{P I_w} I(A_i = d(0)(w^*, v), W_i = (w^*, v)) \right\} p(w^* \mid v) \\
&= \tfrac{1}{n} \textstyle\sum_i \left\{ \frac{Y_i I(V_i = v, A_i = d(0)(W_i))}{P I_{W_i}} - \frac{\theta_0(W_i)}{P I_{W_i}} I(V_i = v, A_i = d(0)(W_i)) \right\} p(W_i^* \mid V_i),
\end{aligned}$$

where $P I_{W_i} = g(d(0)(W_i) \mid W_i) p_W(W_i)$.

Let's now study $(p_n - p)(w^* \mid v)$. We have

$$p_n(w^* \mid v) = \frac{\sum_i I(W_i^* = w^*, V_i = v)}{\sum_i I(V_i = v)} \equiv \frac{P_n I_{w^* v}}{P_n I_v}.$$

Thus, as above,

$$(p_n - p)(w^* \mid v) \approx \frac{P_n I_{w^* v}}{P I_v} - \frac{p(w^* \mid v)}{P I_v} P_n I_v.$$

Thus, the second term in linearization of $\theta_n - \theta_0$ is given by

$$\begin{aligned}
&\tfrac{1}{n} \textstyle\sum_i \sum_{w^*} \theta_0(w^*, v) \left\{ \frac{I(W_i^* = w^*, V_i = v)}{p(v)} - \frac{p(w^* \mid v)}{p(v)} I(V_i = v) \right\} \\
&= \tfrac{1}{n} \textstyle\sum_i \theta_0(W_i) \frac{I(V_i = v)}{p(V_i)} - \tfrac{1}{n} \textstyle\sum_i \sum_{w^*} \theta_0(w^*, V_i) \frac{p(w^* \mid V_i)}{p(V_i)} I(V_i = v).
\end{aligned}$$

We have $\theta_n - \theta_0(v)$ involves $\sum_i I(V_i = v) h$. Thus,

$$P_0 \sum_{a \in \mathcal{A}_1} h(a, V) \frac{d}{d\beta_0} m(a, V \mid \beta_0)(\theta_n(V) - \theta_0(V))$$

equals

$$\begin{array}{l}\frac{1}{n}\sum_i\sum_a h(a,V_i)\frac{d}{d\beta_0}m(a,V_i\mid\beta_0)\\ \left\{\frac{Y_iI(A_i=d(0)(W_i))}{g(d(0)(W_i)|W_i)p(W_i)}-\frac{\theta_0(W_i)}{g(d(0)(W_i)|W_i)p(W_i)}I(A_i=d(0)(W_i))\right\}p(W_i^*\mid V_i)p(V_i)\\ +\frac{1}{n}\sum_i\sum_a h(a,V_i)\frac{d}{d\beta_0}m(a,V_i\mid\beta_0)\left\{\theta_0(W_i)\frac{1}{p(V_i)}-\sum_{w^*}\theta_0(w^*,V_i)\frac{p(w^*|V_i)}{p(V_i)}\right\}p(V_i).\end{array}$$

Thus,

$$IC_{nu} \;=\; \sum_{a\in\mathcal{A}_1} h(a,V_i)\frac{d}{d\beta_0}m(a,V_i\mid\beta_0)\left\{\frac{I(A_i=d(0)(W_i)}{g(d(0)(W_i)\mid W_i)}\left(Y_i-\theta_0(W_i)\right)+\theta_0(W_i)-\theta_0(V_i)\right\}.$$

We have established the following useful lemma:

Lemma 1 *We have for the nonparametric estimator θ_n of θ_0 the following linear expansion*

$$\begin{array}{l}\sum_v h^*(v)(\theta_n-\theta_0)(v)p_0(v)\approx\\ h^*(V_i)\left\{\frac{I(A_i=d(0)(W_i))}{g(d(0)(W_i)|W_i)}\left(Y_i-\theta_0(W_i)\right)+\theta_0(W_i)-\theta_0(V_i)\right\}.\end{array}$$

Recall that IC_{nu} is the correction factor which needs to be subtracted from the original estimation function which was given by

$$\begin{array}{l}\sum_{a\in\mathcal{A}_1} I(A=d(a)(W))\frac{h(a,V)\frac{d}{d\beta_0}m(a,V|\beta_0)}{g_0(A|X)}(Y-Q_{02}(A,W))\\ +\sum_{a\in\mathcal{A}_1} h(a,V)\frac{d}{d\beta_0}m(a,V\mid\beta_0)(Q_{02}(d(a),W)-m(a,V\mid\beta_0)-\theta_0(V))\end{array}$$

This yields the following corrected estimating function:

$$\begin{array}{l}D_h(\beta,Q,g)=\sum_{a\in\mathcal{A}_1} h(a,V)\frac{d}{d\beta_0}m(a,V\mid\beta_0)\\ \left\{\frac{I(A=d(a)(W)}{g(d(a)(W)|W)}-\frac{I(A=d(0)(W)}{g(d(0)(W)|W)}\right\}(Y-Q_{02}(A,W))\\ +\sum_{a\in\mathcal{A}_1} h(a,V)\frac{d}{d\beta_0}m(a,V\mid\beta_0)\left\{Q_{02}(d(a)(W),W)-Q_{02}(d(0)(W),W)-m(a,V\mid\beta_0)\right\}\\ \equiv D_{1h}(Q,g)+D_{2h}(\beta(Q_0),Q_0),\end{array}$$

where D_{1h} is a score of $P(Y\mid A,W)$ and D_{2h} is a score of the marginal distribution of W.

Double robustness of estimating function: It follows that

$$E_0D_h(\beta(Q),Q,g)=0 \text{ if } g=g_0 \text{ or } Q_2=Q_{02},$$

where the unbiasedness for $g=g_0$ relies on $g(d(a)(W)\mid W)g(d(0)(W)\mid W)>0$: i.e. the rules $d(a)$ have to be realistic for $a\in\mathcal{A}_1$.

Identity: The fact that

$$E_Q \sum_{a \in \mathcal{A}_1} h(a,V) \frac{d}{d\beta_0} m(a,V \mid \beta(Q)) \{Q_2(d(a)(W),W) - Q_2(d(0)(W),W) - m(a,V \mid \beta(Q))\} = 0$$

is a direct consequence of the definition of $\beta(Q)$. We observe that we can also define $\beta(Q)$ as a weighted least squares solution

$$\beta(Q) = \arg\min_\beta E_V \sum_a h(a,V) \left(Q_2(d(a)(W),W) - Q_2(d(0)(W),W) - m(a,V \mid \beta)\right)^2.$$

Thus, given an estimator Q_n with Q_{1n} the empirical distribution of $W_1, \ldots, W_n$, it follows that $\beta(Q_n)$ is given by the weighted least squares solution:

$$\beta(Q_n) = \arg\min_\beta \sum_{i=1}^n \sum_a h(a,V_i) \left(Q_{2n}(d(a)(W_i),W_i) - Q_{2n}(d(0)(W_i),W_i) - m(a,V_i \mid \beta)\right)^2.$$

Targeted MLE: Let $\{Q_2(\epsilon) : \epsilon\}$ be a path through Q_2 at $\epsilon = 0$ and satisfy the score condition $\frac{d}{d\epsilon} \log Q_2(\epsilon)\big|_{\epsilon=0} = D_1^*(Q_2, g_0)$. For example, if Q_2 is a regression model of Y on A, W with normal errors with constant variance, then we can simply add the extension $\epsilon C^*(A,W)$, where

$$C^*(A,W) \equiv \sum_{a \in \mathcal{A}_1} h(a,V) \frac{d}{d\beta_0} m(a,V \mid \beta_0) \left\{ \frac{I(A = d(a)(W)}{g(d(a)(W) \mid W)} - \frac{I(A = d(0)(W)}{g(d(0)(W) \mid W)} \right\}.$$

In other words, $E_{Q_2(\epsilon)}(Y \mid A,W) = E_{Q_2}(Y \mid A,W) + \epsilon C^*(A,W)$.

Similarly, if Q_2 is a logistic regression fit of a binary Y on A, W, then we simply add $\epsilon C^*(A,W)$ to the logit of $Q_2(1 \mid A,W)$. In other words,

$$\text{logit} E_{Q_2(\epsilon)}(Y \mid A,W) = \text{logit} E_{Q_2}(Y \mid A,W) + \epsilon C(A,W).$$

In both cases, these ϵ extensions have a score at $\epsilon = 0$ equal to $D_1^*(Q_2, g_0)$.

Making the epsilon-covariate extension independent of β_0: The targeted MLE can be obtained in one ML step in the case that the epsilon-covariate $C(A,W)$ does not depend on β_0. In the case that $m(a,V \mid \beta)$ is a linear regression model, say $m(a,V \mid \beta) = \beta(a,V)$, then $\frac{d}{d\beta_0} m(a,V \mid \beta_0) = (a,V)^\top$ so that indeed the ϵ-covariate is independent of β_0 for each choice of h.

The one-step targeted MLE: Given an estimate g_n of the treatment mechanism g_0, an estimate $Q_{2\theta_n}$, let ϵ_n be the solution of

$$0 = \sum_i D_1^*(Q_{2\theta_n}(\epsilon_n), g_n)(O_i).$$

In the above two linear and logistic regression ϵ-extensions, and under the assumption that the ϵ-extension covariate $C(A,W)$ does not depend on β_0, it follows that

$$\epsilon_n = \arg\max_{\epsilon} \sum_{i=1}^{n} \log Q_{2\theta_n}(\epsilon)(O_i)$$

is the maximum likelihood estimator over ϵ.

We call $\beta_n = \beta(Q_{1n}, Q_{2\theta_n}(\epsilon_n))$ the targeted MLE of β_0. As mentioned above, one can view β_n as a weighted least squares solution of the regression of $Q_{2\theta_n}(\epsilon_n)(d(a),W_i) - Q_{2\theta_n}(d(0),W_i)$ on the realistic MSM $m(a,V_i \mid \beta)$:

$$\beta_n = \arg\min_{\beta} \sum_{a\in\mathcal{A}_1} h(a,V_i)(Q_{2\theta_n}(\epsilon_n)(d(a),W_i) - Q_{2\theta_n}(d(0),W_i) - m(a,V_i \mid \beta))^2.$$

15 Targeted MLE for semi-parametric relative risk MSM for realistic interventions.

Consider the model $E(Y(d(a)) \mid V)/E(Y(d(0)) \mid V) = m(a,V \mid \beta_0)$ for the causal relative risk. The resulting estimator can also be used as V-adjusted multiplicative variable importance. Let $O = (W,A,Y)$.

The first challenge is to determine the orthogonal complement of the nuisance tangent space $T_{nuis}^{\perp}(P)$ at a P in this model, or a rich subset of this space. For this purpose we follow the strategy from the previous section again (Hubbard, van der Laan (2007, Population Intervention models)) to derive the class of estimating functions for such semi-parametric causal models.

Firstly, we note that $E(Y(d(a)) \mid V) = \theta_0(V)m(a,V \mid \beta_0)$, where $\theta_0(V) = E(Y(d(0)) \mid V) = E(E(Y \mid A = d(0)(W),W) \mid V)$. In the MSM-model with θ_0 known, the orthogonal complement of the nuisance tangent space at P_0 is given by

$$\begin{aligned} D_h^*(\beta_0,Q_0,g_0,\theta_0) &= \sum_{a\in\mathcal{A}_1} I(A = d(a)(W)) \frac{h(a,V)\frac{d}{d\beta_0}m(a,V \mid \beta_0)}{g_0(A \mid X)}(Y - Q_{02}(A,W)) \\ &+ \sum_{a\in\mathcal{A}_1} h(a,V)\frac{d}{d\beta_0}m(a,V \mid \beta_0)(Q_{02}(d(a)(W),W) - m(a,V \mid \beta_0)\theta_0(V)) \\ &\equiv D_{1h}^*(\beta_0,Q_0,g_0,\theta_0)(W,A,Y) + D_{2h}^*(\beta_0,Q_0,\theta_0)(W), \end{aligned}$$

where we defined $Q_{02}(d(a),W) = E_{Q_0}(Y \mid A = d(a)(W),W)$ and $Q_{02}(a,W) = E(Y \mid A = a,W)$, and we note that $\beta_0 = \beta(Q_0)$ is a parameter of $Q_0 = (Q_{01},Q_{02})$.

Let θ_n be an estimator of θ_0. Let IC_{nu} be the influence curve of $-P_0 D_h^*(Q_0, g_0, \theta_n) - D_h^*(Q_0, g_0, \theta_0)$. Then $-P_0(D_h^*(\beta_n) - D_h^*(\beta_0)) \approx (P_n - P_0)(D_h^* - IC_{nu})$. Thus, we need to determine the influence curve IC_{nu} for a nonparametric estimator θ_n of θ_0.

We note that

$$\begin{aligned} -P_0(D_h^*(\theta_n) - D_h^*(\theta_0)) &\approx P_0 \sum_a h(a,V) \frac{d}{d\beta_0} m(a,V \mid \beta_0) m(a,V \mid \beta_0)(\theta_n - \theta_0)(V) \\ &\equiv \sum_v h^*(v)(\theta_n - \theta_0)(v) p(v). \end{aligned}$$

where

$$h^*(v) \equiv \sum_{a \in \mathcal{A}_1} h(a,v) \frac{d}{d\beta_0} m(a,v \mid \beta) m(a,v \mid \beta_0).$$

Recall from previous section that

$$\begin{array}{l} \sum_v h^*(v)(\theta_n - \theta_0)(v) p(v) \\ = \frac{1}{n} \sum_i h^*(V_i) \frac{I(A_i = d(0)(W_i))}{g(d(0)(W_i) \mid W_i)} \{Y_i - \theta_0(W_i)\} \\ + \frac{1}{n} \sum_i h^*(V_i) \{\theta_0(W_i) - \theta_0(V_i)\}. \end{array}$$

Thus,

$$IC_{nu} = h^*(V_i) \left\{ \frac{I(A_i = d(0)(W_i)}{g(d(0)(W_i) \mid W_i)} (Y_i - \theta_0(W_i)) + \theta_0(W_i) - \theta_0(V_i) \right\},$$

where $\theta_0(w) = E(Y \mid A = d(0)(w), W = w)$.

Note IC_{nu} is the correction factor which needs to be subtracted from original estimation function D_h^*. Recall $h^*(V) = \sum_a h(a,V) \frac{d}{d\beta} m(a,V \mid \beta) m(a,V \mid \beta_0)$. This gives the following class of estimating functions indexed by an arbitrary function h for our semi-parametric model:

$$\begin{array}{l} D_h(\beta, Q, g) \equiv \sum_{a \in \mathcal{A}_1} \frac{I(A=d(a)(W))}{g(A|X)} h(a,V) \frac{d}{d\beta} m(a,V \mid \beta)(Y - Q_2(A,W)) \\ + \sum_{a \in \mathcal{A}_1} h(a,V) \frac{d}{d\beta} m(a,V \mid \beta_0)(Q_2(d(a)(W),W) - m(a,V \mid \beta)\theta(V)) \\ - \sum_{a \in \mathcal{A}_1} h(a,V) \frac{d}{d\beta} m(a,V \mid \beta) m(a,V \mid \beta) \{ \\ \frac{I(A=d(0)(W))}{g(A|X)} (Y - Q_2(d(0)(W),W)) + Q_2(d(0)(W),W) - \theta(V) \Big\} \\ \sum_{a \in \mathcal{A}_1} h(a,V) \frac{d}{d\beta} m(a,V \mid \beta) \left\{ \frac{I(A=d(a)(W))}{g(A|X)} - m(a,V \mid \beta) \frac{I(A=d(0)(W))}{g(A|X)} \right\} (Y - Q_2(A,W)) \\ + \sum_{a \in \mathcal{A}_1} h(a,V) \frac{d}{d\beta} m(a,V \mid \beta_0)(Q_2(d(a)(W),W) - m(a,V \mid \beta)\theta(V)) \\ - \sum_{a \in \mathcal{A}_1} h(a,V) \frac{d}{d\beta} m(a,V \mid \beta) m(a,V \mid \beta) \{Q_2(d(0)(W),W) - \theta(V)\} \\ \equiv D_{h1}(\beta, Q, g) + D_{h2}(\beta, Q), \end{array}$$

where $\theta = \theta(Q) = E_Q(Q_2(d(0)(W), W) \mid V)$ and $D_{2h} = D_{2ha} + D_{2hb}$ is a sum of two terms itself.

Double robustness: We have the following double robustness result

$$P_0 D_h(\beta_0, Q, g) = 0 \text{ if either } Q = Q_0 \text{ or } g = g_0.$$

The robustness at $Q = Q_0$ is trivially shown. For the robustness at $g = g_0$, one needs to use that $m(a, V \mid \beta_0) = \theta_0(a, V)/\theta_0(v)$, where $\theta_0(a, V) = E_0(Y(d(a)) \mid V) = E_0(Q_{02}(d(0)(W), W) \mid V)$.

The targeted MLE: The ϵ-extension of $Q^0(\epsilon)(Y \mid A, W)$ needs to have score at $\epsilon = 0$ equal to $D^*_{1h}(Q^0, g)$. This can be achieved by adding the following covariate $C(A, W)$ to a logistic regression fit:

$$\sum_{a \in \mathcal{A}_1} h(a, V) \frac{d}{d\beta^0} m(a, V \mid \beta^0) \left\{ \frac{I(A = d(a)(W))}{g(A \mid X)} - m(a, V \mid \beta^0) \frac{I(A = d(0)(W))}{g(A \mid X)} \right\}.$$

Since this covariate will depend on β itself, one will need to iterate the t-mle step till $\epsilon^k_n \approx 0$. At that step we will have an updated Q^k_n so that $P_n D_{h1}(Q^k_n, g_n 0 = o_P(1/\sqrt{n})$. We now compute a corresponding $\theta_n(V) = E(Q^k_{n2}(d(a)(W), W) \mid V)$ w.r.t. to a nonparametrically fit conditional distribution $p(W \mid V)$ of W, given V, so that $P_n D_{h2b}(\beta, Q^k) = 0$ for all β:

$$0 = \sum_i \sum_{a \in \mathcal{A}_1} h(a, V_i) \frac{d}{d\beta} m(a, V_i \mid \beta) m(a, V_i \mid \beta) \left\{ Q^k_2(d(0)(W_i), W_i) - \theta^k(V_i) \right\}.$$

This can always be arranged by estimating $p(W \mid V)$ nonparametrically enough, at most using some smoothing if needed. Given $Q^k_2(d(a)(W), W)$ and θ^k, one now solves for β^k so that

$$0 = \sum_i \sum_{a \in \mathcal{A}_1} h(a, V_i) \frac{d}{d\beta} m(a, V_i \mid \beta)(Q^k_2(d(a)(W_i), W_i) - m(a, V_i \mid \beta^k)\theta^k(V_i)).$$

The latter corresponds with regressing $Q^k_2(d(a)(W_i), W_i)$ on $m(a, V_i \mid \beta^k)\theta^k(V_i)$. In this way, we guarantee that

$$0 = \sum_i D_h(\beta(Q^k), Q^k, g_n)(O_i)$$

so that the targeted MLE $\beta(Q^k)$ can be analyzed as the double robust estimator based on the double robust estimating function $D_h(\beta, Q, g)$.

16 Targeted MLE for V-adjusted additive variable importance for continuous A.

Assume $E(E(Y \mid A=a,W) - E(Y \mid A=0,W) \mid V) = m(a,V \mid \beta_0)$. The above derived t-mle for the semiparametric additive msm for realistic rules based on a working model $m(\cdot \mid \beta)$ (so β_0 is defined nonparametrically in terms of least squares projection) needs to assume that A is discrete since it relies on an estimate of the probability that $A=0$, given covariates. The approach followed there estimated $\theta_n(V)$ without using the working model (i.e., one cannot get $E(Y(0) \mid V)$ by extrapolating from $E(Y(a) \mid V)$), which explains why our resulting class of estimating functions relies on a positive probability on $A=0$. The resulting estimator had a well understood projection extension if the working model is misspecified.

Here we derive a class of double robust estimating functions which apply to general A and we restrict to the effect of static interventions. If the model $m(\cdot \mid \beta)$ is wrong, these estimating functions correspond with particular projections of the true V-adjusted var imp on the (working) model.

The idea is too inverse weight the estimating functions for $E(Y \mid A,V) - E(Y \mid A=0,V) = m(a,V \mid \beta_0)$ under $g^*(A \mid X) = g^*(A \mid V)$ with g^*/g, and subsequently orthogonalize them w.r.t. T_{CAR}. This results in the following class of estimating functions indexed by h and g^*:

$$\begin{aligned}
D_h^*(\beta_0,\theta_0,Q,g) &= \frac{g^*(A \mid V)}{g(A \mid W)} \{h(A,V) - E_{g^*}(h(A,V) \mid V)\} (Y - m(A,V \mid \beta_0) - \theta_0(V)) \\
&-\frac{g^*(A \mid V)}{g(A \mid W)} \{h(A,V) - E_{g^*}(h(A,V) \mid V)\} (Q_{02}(A,W) - m(A,V \mid \beta_0) - \theta_0(V)) \\
&+\sum_a g^*(a \mid V) \{h(a,V) - E_{g^*}(h \mid V)\} (Q_{02}(a,W) - m(a,V \mid \beta_0) - \theta_0(V)) \\
&= \frac{g^*(A \mid V)}{g(A \mid W)} \{h(A,V) - E_{g^*}(h(A,V) \mid V)\} (Y - Q_{02}(A,W)) \\
&+\sum_a g^*(a \mid V) \{h(a,V) - E_{g^*}(h \mid V)\} (Q_{02}(a,W) - m(a,V \mid \beta_0) - \theta_0(V)) \\
&\equiv D_{1h}^*(Q_0,g) + D_{2h}^*(\beta_0,\theta_0,Q)
\end{aligned}$$

This class of estimating functions indexed by h are double robust in the sense that they are solved at the true β_0 if either $g = g_0$ or $Q = Q_0$, (and θ_0 can always be misspecified) and the model is correctly specified.

As h we recommend the choice which yields the efficient influence curve in the case that the data is (V,A,Y) and one assumes the model $E(Y \mid A =$

$a, V) - E(Y \mid A = 0, V) = m(a, V \mid \beta_0)$, as presented above: e.g. in the constant variance case, $h^*(\beta_0)(A, V) = \frac{d}{d\beta_0} m(A, V \mid \beta_0)$. For this choice, let $D_1(\beta, Q, g) = D_{1h(\beta)}(Q, g)$ and $D_2(\beta, \theta, Q) = D_{2h(\beta)}(\beta, \theta, Q)$.

If one works in the nonparametric model, one can view this choice of double robust estimating function as an efficient influence curve in a nonparametric model for a parameter $Q \to \beta(Q)$ defined nonparametrically as a solution of the second component equation: $\beta(Q)$ is the β solving $P_0 D_{2h^*(\beta(Q))}(\beta(Q), \theta(Q), Q) = 0$.

Solving the second component of efficient influence curve equation, given Q^0: We note that, given a Q^0, we can define $(\beta(Q^0), \theta(Q^0))$ as the following iterative targeted MLE type solution. This is important since it provides us with a powerful way to deal with multiple solutions and a fast way to compute the estimator $\beta(Q^0)$. Consider the case that the marginal distribution of W under Q^0 is the empirical probability distribution of W.

Consider an initial estimator β^0, θ^0 (say implied by Q^0) of β_0, θ_0. Now, compute the following weighted linear regression estimator:

$$\epsilon_n^0 \equiv \arg\max_{\epsilon} \sum_i \sum_a g_n^*(a \mid V_i) \left\{ Q^0(a, W_i) - m_{\beta^0 + \epsilon}(a, V_i) - \theta^0(V_i) - \epsilon h^*(\beta^0, g_n^*)(V_i) \right\}^2,$$

where we define $h^*(\beta, g^*) = -E_{g^*}(h^*(\beta)(A, V) \mid V) = -E_{g^*}(\frac{d}{d\beta} m_\beta(A, V) \mid V)$. In the case that $m_\beta(a, V) = \beta a V$, ten $h^*(\beta, g^*) = h^*(g^*)$ does not depend on β. Thus, ϵ_n^0 is obtained as a repeated measures weighted linear least squares regression fit with off-set $\beta^0 aV + \theta^0(V)$, adding covariate extension $\epsilon\{aV + h^*(\beta^0, g_n^*)(V)\}$, where the weights are $g_n^*(a \mid V_i)$. Now, let $\beta^1 = \beta^0 + \epsilon_n^0$ and $\theta^1 = \theta^0 + \epsilon h^*(\beta^0, g_n^*)$. We now iterate this process till convergence (i.e., $\epsilon_n^k \approx 0$) and let $\beta(Q^0) = \beta^k$ and $\theta(Q^0) = \theta^k$ for this large enough choice k. If m_β is linear, then the convergence occurs in a single step so that $\beta_n^k = \beta^1$ and $\theta_n^k = \theta^1$. Because the score of the used ϵ-extension for the final k is solved at $\epsilon = 0$ it follows that $\beta(Q^0), \theta(Q^0)$ solve the estimating equation:

$$\begin{aligned} 0 \;=\; & \sum_i \sum_a g_n^*(a \mid V_i) \left\{ \frac{d}{d\beta(Q^0)} m_{\beta(Q^0)}(a, V_i) + h_1^*(\beta(Q^0), g_n^*)(V_i) \right\} \times \\ & (Q^0(a, W_i) - m_{\beta(Q^0)}(a, V_i) - \theta(Q^0)(V_i)), \end{aligned}$$

or equivalently $\sum_i D_2(\beta(Q^0), \theta(Q^0), Q^0)(O_i) = 0$.

Solving first component of efficient influence curve equation, given β^0, θ^0: Let Q^0 be an initial estimator estimating $Q_{02}(A, W) = E_{Q_0}(Y \mid A, W)$ according to a regression model, and estimating the marginal distribution of W with the empirical distribution of W. Let $Q^0(\epsilon)$ be obtained by adding

$\epsilon C(\beta^0)(A,W)$ to the regression fit $Q_2^0(A,W)$, where

$$C(\beta^0) = \frac{g_n^*(A \mid V)}{g_n(A \mid W)} \left\{ \frac{d}{d\beta^0} m_{\beta^0}(A,V) + h_1^*(\beta^0, g_n^*)(V) \right\}.$$

Let ϵ_n^0 be the MLE and let Q_n^1 be the corresponding update. In principle, we iterate this till $\epsilon_n^k \approx 0$ and thereby Q_n^k solves

$$0 = \sum_i D_1(\beta^0, Q_n^k, g_n)(O_i).$$

However, since the covariate $C(\cdot)$ does not depend on Q_n^k it follows that convergence occurs at the first step so that we already have

$$0 = \sum_i D_1(\beta^0, Q_n^1, g_n)(O_i).$$

Solving the efficient influence curve equation: Given Q^0 and (in non linear case) β^0, we find the one step update Q^1 solving

$$0 = \sum_i D_1(\beta^1, Q^1, g_n)(O_i) = 0.$$

If the model m_β is linear, then D_1 does not depend on β so that we have

$$0 = \sum_i D_1(Q^1, g_n)(O_i) = 0.$$

Given Q^1 (and say corresponding β^0, θ^0), we solve for the iteratively obtained (but in the linear case, single step) update $\beta^1 = \beta(Q^0), \theta^1 = \theta(Q^0)$ solving

$$0 = \sum_i D_2(\beta^1, \theta^1, Q^1)(O_i) = 0.$$

In the linear model case, we are now done since we have

$$0 = \sum_i D_2(\beta^1, \theta^1, Q^1)(O_i) = \sum_i D_1(Q^1, g_n)(O_i)$$

so that the double robust estimating equation is already solved. In general, at each step the log likelihood of Q^k increases in k and we stop till convergence is established (i.e., $\epsilon_n^k \approx 0$ in the update step for Q^k) at which point both estimating equations are solved so that

$$0 = \sum_i D_{DR}(\beta^k = \beta(Q^k), \theta^k = \theta(Q^k), Q^k, g_n)(O_i) = 0.$$

That is, we solved the double robust estimating equation.

For statistical inference we can use that the t-mle $\beta_n = \beta(Q^k)$ solves the double robust estimating equation

$$0 = \sum_i D^*_{h^*(\beta_n),g^*_n}(\beta_n, Q^k, g_n)(O_i).$$

17 Semi-parametric logistic regression and corresponding odds-ratio variable importance.

Let $O = (W, A, Y) \sim P_0$. Assume

$$Q_0(A, W) \equiv P_0(Y = 1 \mid A, W) = \frac{1}{1 + \exp(-\{A\beta_0 W + r_0(W)\})}$$

for some β_0 and function r_0. We wish to construct the iterative targeted MLE of β_0 based on an i.i.d. sample $O_1, \ldots, O_n$ from P_0.

Firstly, we are concerned with construction of the nuisance tangent space of the unspecified g_0 so that we can find the efficient influence curve and corresponding hardest sub-model through a current fit, as needed to define the targeted MLE. For that purpose, we can consider ϵ-paths $P_\epsilon(Y = 1 \mid A, W) = \frac{1}{1+\exp(-\{A\beta_0 W + r_0(W) + \epsilon h(W)\})}$ for arbitrary functions h. This results in the nuisance tangent space

$$T_{nuis,r_0}(P_0) = \{h(W)(Y - Q_0(A, W)) : h\}.$$

We wish to construct a path $Q(\epsilon)$ through Q at $\epsilon = 0$ so that its score at $\epsilon = 0$ is orthogonal to the nuisance tangent space. Since any score is already orthogonal to the nuisance scores generated by the distribution of (A, W), it follows that is suffices to establish that this score is orthogonal to $T_{nuis,r_0}(P_0)$. Consider the candidate paths

$$Q_{0h_1}(\epsilon)(Y = 1 \mid A, W) = \frac{1}{1 + \exp(-\{A(\beta_0 + \epsilon)W + r_0(W) + \epsilon h_1(W)\})}.$$

The score of this path at $\epsilon = 0$ equals

$$S(h_1) \equiv (AW + h_1(W))(Y - Q_0(A, W)).$$

We now need to select h_1 so that for each $h(W)$ we have

$$\begin{aligned} 0 &= E_0(AW + h_1(W))(Y - Q_0(A, W))h(W)(Y - Q_0(A, W)) \\ &= E_0(AW + h_1(W))h(W)\sigma_0^2(A, W)), \end{aligned}$$

where $\sigma_0^2(A,W) = Q_0(A,W)(1-Q_0(A,W))$. It follows that the unique solution is given by

$$h_1^*(Q_0,g_0)(W) = -\frac{E_0\{AWQ_0(1-Q_0)(A,W) \mid W\}}{E_0\{Q_0(1-Q_0)(A,W) \mid W\}},$$

where g_0 denotes the conditional distribution of A, given W. In particular, this shows that the efficient influence curve is up till a scaling matrix given by:

$$D^*(Q_0,g_0)(O) = \{AW + h_1^*(Q_0,g_0)(W)\}(Y - Q_0(A,W)).$$

We note that one can also represent D^* as function in g_0, β_0, r_0:

$$D^*(\beta_0,r_0,g_0)(O) = \{AW + h_1^*(\beta_0,r_0,g_0)(W)\}(Y - Q_{\beta_0,r_0}(A,W)).$$

We are now ready to define the targeted MLE. Let Q^0, g^0 be initial estimators of Q_0, g_0, where Q^0 is defined by (β^0, r^0). Construct path $Q^0_{h_1^*(Q^0,g^0)}(\epsilon)$ and compute MLE ϵ_n^0 of ϵ. This corresponds with fitting a logistic regression model in covariate AW and h_1^*, with offset $\beta^0 AW + r^0(W)$. We now update $Q^1 = Q^0_{h_1^*(Q^0,g^0)}(\epsilon_n^0)$. We iterate this process till $\epsilon_n^k \approx 0$ at which point we have

$$0 = \sum_{i=1}^n D^*(\beta_n^k, r_n^k, g_n^0)(O_i)$$

up till a user supplied numerical precision. The estimator β_n^k's influence curve and thereby statistical inference can now be derived from the fact that it solves this estimating equation.

18 Double robust targeted MLE for direct effect model.

Consider a direct effect model $E(\sum_z Y_{az} Q_0(z \mid W) \mid V) = m(a, V \mid \beta_0)$ for a given distribution $Q_0(z \mid W)$, where one particular choice is the distribution of Z, given $A = 0, W$, typically replaced by an estimator. Let $O = (W, A, Z, Y = Y(A,Z))$ be a missing data structure on $X^* = (W, (Y(a,z) : a, z))$ and assume the usual coarsening at random assumption. Let $g(a, z \mid X^*)$ be the conditional probability distribution of (A, Z), given X^*, which only depends on W (by CAR). In addition, let $Q_{Y0} = E_0(Y \mid A, Z, W)$ and Q_{W0} denotes the true probability distribution of W.

In van der Laan, Petersen (2008), we discussed estimators as solutions of the double robust IPCW estimating equations. An important issue with defining an estimator of β_0 as a solution of estimating equations is that its solutions are typically not compatible with a particular probability distribution of the data, and that there is a lack of criteria to select among a possible set of solutions. Targeted maximum likelihood estimation address these issues by providing a maximum likelihood based estimation procedure resulting in an estimator $Q_n = (Q_{Wn}, Q_{Yn})$ in which its substitution estimator $\beta_n = \beta(Q_n)$ solves the wished double robust estimating equation.

We start with noting that the class of double robust estimating functions indexed by h can be decomposed as a sum of two estimating functions:

$$\begin{aligned}
&D^*_{h,DR}(O \mid \beta_0, g_0, Q_{Y0}) = \\
&\tfrac{g^*(A|V)}{g_0(A,Z|X^*)} \{h_1(A,V) - E_{g^*}(h_1(A,V) \mid V)\} Q_0(Z \mid W)(Y - Q_{Y0}(A,Z,W)) \\
&+ \textstyle\sum_{a,z} g^*(a \mid V) \{h_1(a,V) - E_{g^*}(h_1(A,V) \mid V)\} Q_0(z \mid W) \times \\
&\{Q_{Y0}(a,z,W) - m(a,V \mid \beta_0) - h_2(V)\} \\
&\equiv D^*_{1h}(g_0, Q_{Y0}) + D^*_{2h}(\beta_0, Q_{Y0}),
\end{aligned}$$

where we suppress the dependence on $Q_0(Z \mid W)$.

This class of estimating functions indexed by h are double robust in the sense that they are solved at the true β_0 if either $g = g_0$ or $Q = Q_0$, and the model is correctly specified.

Based on efficiency considerations, as target choice $h = (g^*, h_1, h_2)$ we recommended $h_1(a,V) = \frac{d}{d\beta_0} m(a,V \mid \beta_0)$, $h_2(V) = m_0(V) \equiv E(\sum_z Q_0(z \mid W) Q_{Y0}(0,z,W) \mid V)$, and $g^*(A \mid V) = g_0(A \mid V)$. If $m(\cdot \mid \beta)$ is linear in β, then h_1 is known. Either way, we replace each of these target choices by estimates of the corresponding quantities, resulting in a choice h_n with an asymptotic limit h_∞, not necessarily equal to this wished choice h. We note that under the assumption that $m(a,V \mid \beta_0)$ is correctly specified, $E(\sum_z Y_{az} Q)(z \mid W) \mid V) = m(a,V \mid \beta_0) + m_0(V)$ so that one can view $m(a,V \mid \beta_0) + m_0(V)$ as a semi-parametric additive causal regression model modeling direct effects.

Since it is hard to construct an estimator Q_{Yn} of Q_{Y0} satisfying a particular direct effect model $m(\cdot \mid \beta)$, we wish to work in the nonparametric model in which $m(\cdot\beta)$ is merely viewed as a working model. In this nonparametric model one can view the (standardized version of the) recommended choice of double robust estimating function as an efficient influence curve in a nonparametric model for a parameter $Q = (Q_W, Q_Y) \to \beta(Q)$ defined nonparametrically as a solution of the second component equation: $\beta(Q)$ is a β solving $E_{Q_W} D_{2h^*}(\beta, Q_Y) = 0$, where Q_W denotes a distribution of W and E_{Q_W} the expectation w.r.t. distribution Q_W.

Our particular choice of mapping from a distribution Q_W of W and the conditional mean Q_Y into $\beta(Q_W, Q_Y)$ solving this equation is defined now. We note that if the direct effect model $m(\cdot \mid \beta)$ is correctly specified, then $\beta(Q_0)$ corresponds with the true parameter in this direct effect model.

Nonparametric definition of $\beta(Q)$ solving the second component of efficient influence curve equation: We note that, given an estimator $Q = (Q_W, Q_Y)$, we can define $(\beta(Q), m(Q))$ as follows. We consider the case that the marginal distribution of W under Q is the empirical probability distribution of W.

This definition relies on the specification of an initial estimator β^0, m^0 (which could be implied by $Q = (Q_W, Q_Y)$ itself) of β_0, m_0. Firstly, we compute the following weighted linear regression estimator:

$$\epsilon_n^0 \equiv \arg\max_{\epsilon} \sum_i \sum_{a,z} g_n^*(a \mid V_i) Q_0(z \mid W) \{Q_Y(a, z, W_i) - m_{\beta^0+\epsilon}(a, V_i) - m^0(V_i) - \epsilon h_n^*(V_i)\}^2,$$

where we define $h_n^*(V) = -E_{g_n^*}(h_1(A, V) \mid V) = -E_{g_n^*}(\frac{d}{d\beta^0} m_{\beta^0}(A, V) \mid V)$. Thus, for example, if $m(a, V \mid \beta) = \beta aV$ is linear, then ϵ_n^0 is obtained as a repeated measures weighted linear least squares regression fit with off-set $\beta^0 aV + m^0(V)$, adding covariate extension $\epsilon\{aV + h^*(V)\}$, where the weights are $g*(a \mid V_i)Q_0(z \mid W_i)$.

Let $\beta^1 = \beta^0 + \epsilon_n^0$ and $m^1 = m^0 + \epsilon h^*$ be the updates, which can be interpreted as a targeted estimator of β_0, m_0 targeting β_0, given an initial estimator (β^0, m^0).

In general, if one wishes to update h^* itself as well based on the newly obtained estimate β^1, then, we can iterate this process till convergence (i.e., $\epsilon_n^k \approx 0$) and let $\beta(Q) = \beta^k$ and $m(Q) = m^k$ for this large enough choice k. If either m_β is linear or one simply uses a fixed h^* according to its initial estimate, then the convergence occurs in a single step so that $\beta_n^k = \beta^1$ and $m_n^k = m^1$. For simplicity, and since it comes without cost of asymptotic performance, we recommend to use a fixed h^* so that the first step β^1 already represents the evaluation $\beta(Q)$.

Because the score of the used ϵ-extension for the final k is solved at $\epsilon = 0$ it follows that $\beta(Q), m(Q)$ solve the estimating equation:

$$0 = \sum_i \sum_{a,z} g_n^*(a \mid V_i) Q_0(z \mid W_i) \left\{ \frac{d}{d\beta_n^0} m_{\beta_n^0}(a, V_i) + h_n^*(V_i) \right\} \times (Q^0(a, z, W_i) - m_{\beta(Q)}(a, V_i) - m(Q)(V_i)),$$

or equivalently $\sum_i D_{2h_n}(\beta(Q_W, Q_Y), Q_Y)(O_i) = 0$ with $h_n = (g_n^*, h_n^*, h_2 = m(Q))$.

Solving first component of efficient influence curve equation with targeted MLE: Let $Q^0 = (Q_Y^0, Q_W^0)$ be an initial estimator estimating $Q_{Y0}(A, Z, W) = E_0(Y \mid A, Z, W)$ according to a regression model, and estimating the marginal distribution of W with the empirical distribution of W. Targeted maximum likelihood estimation involves updating such an initial estimator by maximizing the likelihood in a particular direction targeting the parameter $\beta(Q)$ of interest, thereby reducing the bias of this initial likelihood based estimator $\beta(Q^0)$.

Let $Q^0(\epsilon)$ be obtained by adding $\epsilon C_n(A, Z, W)$ to the regression fit $Q_Y^0(A, W)$, while keeping Q_W^0 the same, where

$$C_n(A, Z, W) = \frac{g_n^*(A \mid V)}{g_n(A, Z \mid X^*)} Q_0(Z \mid W) \left\{ \frac{d}{d\beta^0} m_{\beta^0}(A, V) + h_n^*(V) \right\}.$$

Let ϵ_n^0 be the MLE according to a normal regression model:

$$\epsilon_n^0 = \arg\min_{\epsilon} \sum_{i=1}^{n} (Y_i - Q_Y(\epsilon)(A_i, Z_i, W_i))^2.$$

Let Q_n^1 be the corresponding update. In principle, if one decides to update C_n at each step, we iterate this till $\epsilon_n^k \approx 0$ and thereby Q_n^k solves

$$0 = \sum_i D_{1h_n}(Q_n^k, g_n)(O_i).$$

However, if the covariate C_n does not depend on Q_n^k (i.e., we fix it at its initial estimate), then it follows that convergence occurs at the first step so that we already have

$$0 = \sum_i D_{1h_n}(Q_n^1, g_n)(O_i)$$

for $h_n = (g_n^*, h_n^*, h_2)$ for arbitrary choice h_2 (since D_{1h} does not depend on h_2). Again, we recommend to use a fixed C_n so that the targeted MLE is attained in a single step.

The targeted MLE solves the efficient influence curve equation: We start out with specifying an estimator g_n^*, corresponding estimate h_n^*, and an estimator $h_{2n}(V) = m_n^0(V)$ of $\theta_0(V)$, resulting in $h_n = (g_n^*, h_n^*, h_{2n})$. In addition, we specify an initial estimator $Q_n^0 = (Q_{Wn}^0, Q_{Yn}^0)$ of (Q_{W0}, Q_{Y0}),

where we set Q^0_{Wn} equal to the empirical distribution of $W_1, \ldots, W_n$. Given Q^0_n, we find the one step targeted MLE $Q^1_n = Q^0_n(\epsilon^0_n)$ (defined above) solving

$$0 = \sum_i D_{1,g^*_n,h^*_n}(Q^1_n, g_n)(O_i) = 0.$$

Given Q^1_n (and say corresponding β^0_n, m^0_n), we determine $\beta^1_n = \beta(Q^1_n), \theta^1_n = \theta(Q^1_n)$ defined above as a targeted semi-parametric repeated measures regression solving

$$0 = \sum_i D_{2,g^*_n,h^*_n,m(Q^1_n)}(\beta^1, Q^1_n)(O_i) = 0.$$

We conclude that the targeted MLE $\beta^1_n = \beta(Q^1_n)$ solves the double robust estimating equation

$$0 = \sum_i D_{h_n,DR}(\beta^1 = \beta(Q^1_n), Q^1_n, g_n)(O_i) = 0,$$

where $h_n = (g^*_n, h^*_n, \theta(Q^1_n))$.

Since β^1_n solves the double robust estimating equation, statistical inference for β_0 based on the targeted MLE proceeds in the same manner as for the DR-IPTW estimator.

19 Semi-parametric logistic MSM for point treatment.

Let $X = (W, (Y(a) : a \in \mathcal{A}))$ and $O = (A, W, Y = Y(A))$. Consider the working model

$$E(Y(a) \mid V) = m_{\beta_0,r_0}(a, V) \equiv \frac{1}{1 + \exp\left(-\{\beta_0 aV + r_0(V)\}\right)},$$

for some vector β_0 and function r_0, where $V \subset W$. Let $g_0(A \mid X) = g_0(A \mid W)$ be the treatment mechanism satisfying the missing at random assumption. In the special case that $g_0(A \mid X) = g_0(A \mid V)$, the previous section determines an appropriate estimating function:

$$D^*(\beta_0, r_0, g^*_0)(O) = \{AV + h^*_1(\beta_0, r_0, g^*_0)(V)\}\,(Y - m_{\beta_0,r_0}(A, V)),$$

where

$$h^*_1(\beta_0, r_0, g_0)(W) = -\frac{E_0\{AV m_{\beta_0,r_0}(1 - m_{\beta_0,r_0})(A, V) \mid V\}}{E_0\{m_{\beta_0,r_0}(1 - m_{\beta_0,r_0})(A, V) \mid V\}}.$$

Note that D^* equals the efficient influence curve for the semi-parametric logistic regression model m_{β_0,r_0} for Y, given A, V, based on the reduced data $(V, A, Y) \sim P_{\beta_0,r_0,g_0^*}$. The Inverse Probability of Treatment Weighted (IPTW)-version of D^* is given by:

$$D_{IPTW,g_0^*}(\beta_0, r_0, g_0) = \frac{g_0^*(A \mid V)}{g_0(A \mid W)} \{AV + h_1^*(\beta_0, r_0, g_0^*)(V)\} (Y - m_{\beta_0,r_0}(A, V))$$

The IPTW estimator as an iterative targeted MLE procedure: Before we proceed we wish to show that we can represent and compute the corresponding IPTW estimator of β_0 and r_0 as an iterative targeted MLE type estimator. This is important since it provides us with a powerful way to deal with multiple solutions and a fast way to compute the IPTW-estimator. Consider an initial estimator β^0, r^0 of β_0, r_0. Now, compute the following weighted logistic regression estimator:

$$\epsilon_n^0 \equiv \arg\max_{\epsilon} \sum_i \frac{g_n^*(A_i \mid V_i)}{g_n(A_i \mid W_i)} \log$$
$$\left\{ m_{\beta^0+\epsilon,r^0+\epsilon h_1^*(\beta^0,r^0,g_n^*)}(A_i, V_i)^{Y_i} \left(1 - m_{\beta^0+\epsilon,r^0+\epsilon h_1^*(\beta^0,r^0,g_n^*)}(A_i, V_i)\right)^{1-Y_i} \right\}.$$

Thus, ϵ_n^0 is obtained as a weighted logistic regression maximum likelihood fit with off-set $\beta^0 aV + r^0(V)$, adding covariate extension $\epsilon(aV + h_1^*(\beta^0, r^0, g_n^*)(V))$, where the weights are $g_n^*(A_i \mid V_i)/g_n(A_i \mid W_i)$. Now, we update $\beta^1 = \beta^0 + \epsilon_n^0$ and $r^1 = r^0 + \epsilon h_1^*(\beta^0, r^0, g_0^*)$, and iterate this updating process till convergence (i.e., $\epsilon_n^k \approx 0$) and let $\beta_n = \beta^k$ and $r_n = r^k$ for this large enough choice k. Because the score of the used ϵ-extension for the final k is solved at $\epsilon = 0$ it follows that the final updateβ_n, r_n solves the estimating equation:

$$0 = \sum_i \frac{g_n^*(A_i \mid V_i)}{g_n(A_i \mid W_i)} \{A_i V_i + h_1^*(\beta_n, r_n, g_n^*)(V_i)\} (Y_i - m_{\beta_n,r_n}(A_i, V_i)).$$

Thus, β_n solves the IPTW estimator function with nuisance parameter r_0 estimated with the estimator r_n. We can use the weighted log likelihood as criteria to select among different choices of initial estimators r^0, β^0 (e.g. using likelihood based cross-validation).

Double robust estimating function: The with D_{IPTW} corresponding double robust estimating function is obtained by subtracting from D_{IPTW} its projection on all functions of A, W with conditional mean zero, given W, and is thus given by:

$$
\begin{aligned}
D_{DR,g_0^*}(\beta(Q_0), r(Q_0), Q_0, g_0) &= \frac{g_0^*(A \mid V)}{g_0(A \mid W)} \{AV + h_1^*(\beta(Q_0), r(Q_0), g_0^*)(V)\} \\
&\quad \times (Y - Q_0(A, W)) \\
&+ \sum_a g_0^*(a \mid V) \{aV + h_1^*(\beta(Q_0), r(Q_0), g_0^*)(V)\} \\
&\quad \times (Q_0(a, W) - m_{\beta(Q_0), r(Q_0)}(a, V)) \\
&\equiv D_{1g_0^*}(\beta(Q_0), r(Q_0), Q_0, g_0) \\
&\quad + D_{2g_0^*}(\beta(Q_0), r(Q_0), Q_0, g_0),
\end{aligned}
$$

where Q_0 represents the conditional distribution of Y, given A, W, and the marginal distribution of W, and, given Q_0, we define $\beta(Q_0), r(Q_0)$ so that it is a solution of $P_0 D_{2g_0^*}(\beta, r, Q_0, g_0) = 0$.

Solving the second component of efficient influence curve equation, given Q^0: We note that, given a Q^0, we can define $(\beta(Q^0), r(Q^0))$ as the following iterative targeted MLE type solution. This is important since it provides us with a powerful way to deal with multiple solutions of the efficient influence curve equation, and a fast way to compute the estimator $\beta(Q^0)$. Consider the case that the marginal distribution of W under Q^0 is the empirical probability distribution of W.

Consider an initial estimator β^0, r^0 of β_0, r_0. Now, compute the following weighted logistic regression estimator:

$$
\epsilon_n^0 \equiv \arg\max_{\epsilon} \sum_i \sum_a \frac{g_0^*(a \mid V_i)}{m_{\beta^0, r^0}(1 - m_{\beta^0, r^0})(a, V_i)} \left\{ Q^0(a, W_i) - m_{\beta^0 + \epsilon, r^0 + \epsilon h_1^*(\beta^0, r^0, g_0^*)}(a, V_i) \right\}^2.
$$

Thus, ϵ_n^0 is obtained as a repeated measures weighted least squares logistic regression fit with off-set $\beta^0 aV + r^0(V)$, adding covariate extension $\epsilon(aV + h_1^*(\beta^0, r^0, g_0^*)(V))$, where the weights are $g_n^*(a \mid V_i)/m_{\beta^0, r^0}(1 - m_{\beta^0, r^0})(a \mid V_i)$. Now, let $\beta^1 = \beta^0 + \epsilon_n^0$ and $r^1 = r^0 + \epsilon h_1^*(\beta^0, r^0, g_0^*)$. We now iterate this process till convergence (i.e., $\epsilon_n^k \approx 0$) and let $\beta(Q^0) = \beta^k$ and $r(Q^0) = r^k$ for this large enough choice k. Because the score of the used ϵ-extension for the final k is solved at $\epsilon = 0$ it follows that $\beta(Q^0), r(Q^0)$ solve the estimating equation:

$$
0 = \sum_i \sum_a g_n^*(a \mid V_i) \left\{ aV_i + h_1^*(\beta(Q^0), r(Q^0), g_n^*)(V_i) \right\} (Q^0(a, W_i) - m_{\beta(Q^0), r(Q^0)}(a, V_i)),
$$

or equivalently $\sum_i D_{2g_n^*}(\beta(Q^0), r(Q^0), Q^0)(O_i) = 0$.

Solving first component of efficient influence curve equation, given β^0, r^0: Let Q^0 be an initial estimator estimating $Q_0(A, W) = P_0(Y = 1 \mid A, W)$

according to a logistic regression model and estimating the marginal distribution of W with the empirical distribution of W. Let $Q^0(\epsilon)$ be obtained by adding $\epsilon C(\beta^0, r^0)(A, W)$ to the logit of $Q^0(Y \mid A, W)$, where

$$C(\beta^0, r^0) = \frac{g_n^*(A \mid V)}{g_n(A \mid W)} \left\{ AV + h_1^*(\beta^0, r^0, g_n^*)(V) \right\}.$$

Let ϵ_n^0 be the MLE and let Q_n^1 be the corresponding update. In principle, we iterate this till $\epsilon_n^k \approx 0$ and thereby Q_n^k solves

$$0 = \sum_i D_{1g^{0*}}(\beta^0, r^0, Q_n^k, g_n^0)(O_i).$$

However, since the covariate does not depend on Q_n^k it follows that convergence occurs at the first step so that we already have

$$0 = \sum_i D_{1g^{0*}}(\beta^0, r^0, Q_n^1, g_n^0)(O_i).$$

Solving the efficient influence curve equation: Given Q^0 (and say corresponding β^0, r^0), we solve for the iteratively obtained update $\beta^1 = \beta(Q^0), r^1 = r(Q^0)$ solving

$$0 = \sum_i D_{2g_n^*}(\beta^1, r^1, Q^0)(O_i) = 0.$$

Given β^1, r^1, we find one step update Q^1 solving

$$0 = \sum_i D_{1g_n^*}(\beta^1, r^1, Q^1, g_n)(O_i) = 0.$$

At each step the log likelihood of Q^k increases in k and we stop till convergence is established (i.e., $\epsilon_n^k \approx 0$ in the update step for Q^k) at which point both estimating equations are solved so that

$$0 = \sum_i D_{DR,g^{*0}}(\beta^k = \beta(Q^k), r^k = r(Q^k), Q^k, g_n)(O_i) = 0.$$

That is, we solved the double robust estimating equation.

The influence curve of β_n^k can now be derived from the fact that it solves this estimating equation and statistical inference can be based on this influence curve.

20 IPCW-Reduced Data Targeted MLE.

Let $X = (L_a : a = (a(0), \ldots, a(K)) \in \mathcal{A})$ be a collection of action specific random variables L_a indexed by action regimen a, and let the observed data structure be given by

$$O = (A, L = L_A) = (L(0), A(0), \ldots, L_A(K), A(K), L_A(K+1).$$

The latter represents the time ordering which implies that $L_a(t) = L_{\bar{a}(t-1)}(t)$. Typically, $L_a(t)$ includes a component $R_a(t) = I(T_a \leq t)$ for a failure/end of follow up time T_a, and $L_a(t) = L_a(\min(t, T_a))$ becomes degenerate after the counterfactual time variable T_a. The action process $A(t)$ can have various components describing censoring as well as treatment actions at time t, and for certain values of $A(t-1)$, such as values implying right-censoring, the future process $A(t), \ldots, A(K)$ will be a deterministic function of $\bar{A}(t-1) = (A(0), \ldots, A(t-1))$. In addition, typically, certain values of the observed history $\bar{L}(t), \bar{A}(t-1)$, such as one implying the failure time event $T_A = t$, will determine the future values $A(t), \ldots, A(K)$.

We assume the sequential randomization assumption on the conditional distribution of A, given X, which implies the coarsening at random assumption:

$$g(a \mid X) = \prod_t g_t(a(t) \mid \bar{A}(t-1) = \bar{a}(t-1), X)$$
$$\stackrel{SRA}{=} \prod_t g_t(a(t) \mid \bar{A}(t-1) = \bar{a}(t-1), L_{\bar{A}(t-1)}(t)),$$

where, by support restrictions on $\mathcal{A}$ and the possibly deterministic relation between an observed history $\bar{A}(t-1), \bar{L}(t)$ and the future action process $A(t), \ldots, A(K)$, this product over time t can often be represented as

$$g_0(a \mid X) = \prod_{t=0}^{\min(T_a-1, C_a)} g_t(a(t) \mid \bar{A}(t-1) = \bar{a}(t-1), \bar{L}(t)) \prod_{t=\min(T_a-1, C_a)+1}^{K} I(a(t+1) = a(t)),$$

where C_a denotes the censoring/end of follow up time implied by action regimen a.

Under this CAR/SRA, the probability distribution of the observed data random variable $O = (A, L_A)$ for a single experimental unit factorizes in a

factor Q_0 implied by the full data distribution of X and a factor $g_0(\cdot \mid X)$.

$$\begin{aligned} dP_{Q_0,g_0}(O) &= \prod_{t=0}^{K+1} P_{Q_0}(L(t) \mid \bar{L}(t-1), \bar{A}(t-1)) g_0(A \mid X) \\ &\equiv \prod_{t=0}^{K+1} Q_{0t}(L(t) \mid \bar{L}(t-1), \bar{A}(t-1)) g_0(A \mid X), \end{aligned}$$

where, by CAR we have $Q_{0t}(l(t) \mid \bar{l}(t-1), \bar{a}(t-1)) = P(L_a(t) = l(t) \mid \bar{L}_a(t-1) = \bar{l}(t-1))$ so that indeed Q_0 represents the identifiable part of the full data distribution of X.

Consider a particular model $\mathcal{M} = \{P_{Q,g_0} : Q \in \mathcal{Q}, g_0 \in \mathcal{G}_1\}$ implied by a model $\mathcal{Q}$ for Q_0 and a model $\mathcal{G}_1$ for the censoring mechanism g_0 contained in the set $\mathcal{G}$ of all SRA-conditional distributions of A, given X. Consider also a particular parameter $\Psi : \mathcal{Q} \to \mathbb{R}^d$ defined on this model $\mathcal{Q}$ for Q_0, and let $\psi_0 = \Psi(Q_0)$ denote the true parameter value. Since Q_0 is identifiable, one can also view Ψ as a parameter on the model $\mathcal{M}$ of possible data generating distributions of O.

In this article we provide a class of so called Inverse Probability of Censoring Weighted-Reduced Data- Targeted Maximum Likelihood estimator (IPCW-R-TMLE), obtained by applying the iterative targeted MLE for a reduced data structure but using inverse probability of censoring weighted log-likelihoods at each step. The general targeted MLE methodology is proposed and developed in **?** and can thus also be applied to the complete longitudinal data structure O, as illustrated earlier. The advantage of the IPCW-R-TML estimators is mainly of a practical nature. That is, the IPCW-R-TMLE is often far less complex (and thereby much easier to implement with standard software packages implementing maximum likelihood procedures for the reduced data) than the actual targeted MLE for the actual observed longitudinal data structure which includes time-dependent covariate processes, while the IPCW-R-TMLE still preserves and improves upon important efficiency and robustness properties of the targeted MLE for the reduced data structure. Specifically, an IPCW-R-TML estimator is defined by the following steps.

Specify Reduced Data Structure: Determine a reduction $O^r = (A, L_A^r)$ (i.e., O^r is a function of O), where L_A^r is a measurable function of L_A, where the reduction needs to be so that it is still possible to identify the parameter of interest ψ_0 from the probability distribution of O_r under the under the SRA assumption for the reduced full data structure $X^r = (L_a^r : a \in \mathcal{A})$. For example, $O = (W = L(0), A, \bar{L}(K), Y = L(K+1))$ consists

of baseline covariates W, treatment regimen $A = (A(0), \ldots, A(K))$, time dependent covariate process $\bar{L}(K)$, and a final outcome Y, while one defines $O^r = (W, A, Y)$, which is obtained from O by deleting all time-dependent covariates.

Reduced Data Model. Consider the corresponding reduced data SRA model $\mathcal{M}^r = \{P^r_{Q^r,g^r} = Q^r g^r : Q^r \in \mathcal{Q}^r, g^r \in \mathcal{G}^r\}$ (as described above in general) for $O^r = (A, L^r_A)$, where $\mathcal{G}^r$ is a set of conditional distributions of A, given $X^r = (L^r_a : a \in \mathcal{A})$, satisfying the SRA assumption for the reduced data structure O^r, and $\mathcal{Q}^r$ is a model for the identified component Q^r_0 of the full data distribution of X^r: since Q^r_0 is a function of Q_0, it follows that the model $\mathcal{Q}^r = \{Q^r : Q \in \mathcal{Q}\}$ for Q^r_0 is implied by model $\mathcal{Q}$ for Q_0. Let $\Psi^r : \mathcal{Q}^r \to \mathbb{R}^d$ be such that $\Psi^r(Q^r) = \Psi(Q)$ for all $Q \in \mathcal{Q}^r$, and, in particular, $\Psi^r(Q^r_0) = \Psi(Q_0)$.

Factorization of Q^r: Suppose $dP_{Q^r_0,g^r_0} = \prod_j Q^r_{j0} g^r_0$ factors in various terms Q^r_{j0}, $j = 1, \ldots, J$ (e.g., $J = K + 1$). Suppose that $Q^r_{j0}(O^r)$ depends on O^r only through $((A(0), \ldots, A(j^r - 1), \bar{L}^r(j^r))$, $j = 1, \ldots, J$. In a typical scenario, we have that Q^r_{j0} denotes the conditional distribution of $L^r(j^r)$, given $(A(0), \ldots, A(j^r - 1)$ and $\bar{L}^r(j^r - 1)$. For notational convenience, we used the short-hand notation $j^r = j^r(j)$.

Determine Q^r_j-components of efficient influence curve for reduced data model: Let $D^r(P^r)$ be the efficient influence curve at $dP^r = dP^r_{Q^r,g^r} = Q^r g^r$ for the parameter Ψ^r in the model $\mathcal{M}^r$. This efficient influence curve can be decomposed as:

$$D^r(P^r) = D^r(Q^r, g^r) = \sum_{j=1}^{J} D^r_j(P^r),$$

where $D^r_j(P^r)$ is an element of the tangent space generated by the j-th factor Q^r_j of $Q^r = \prod_j Q^r_j$ at P^r.

Determine hardest Q^r_j-fluctuation functions: Given a Q^r construct submodels $\{Q^r_j(\epsilon) : \epsilon\}$ through Q^r_j at $\epsilon = 0$, with score at $\epsilon = 0$ equal to $D^r_j(Q^r, g^r)$:

$$\frac{d}{d\epsilon} \log Q^r_j(\epsilon)\Big|_{\epsilon=0} = D^r_j(Q^r, g^r), \ j = 1, \ldots, J.$$

Construct IPCW-weights for each Q_j^r-factor: For each j construct weight-function

$$w_j = \frac{g^r(\bar{A}(j^r) \mid X^r)}{g_0(\bar{A}(j^r) \mid X)},\ j = 1, \ldots, J.$$

In short, we will often represent the weights $g^r(\bar{A}(j^r) \mid X^r)/g_0(\bar{A}(j^r) \mid X)$ as g_j^r/g_{0j}. We note

$$\begin{aligned} Q_{j0}^r &= \arg\max_{Q_j^r \in \mathcal{Q}_j^r} P_{Q_0,g_0}\{\log Q_j^r w_j\} \\ &= \arg\max_{Q_j^r \in \mathcal{Q}_j^r} P_{Q_0^r,g_0^r} \log Q_j^r,\ j = 1, \ldots, J. \end{aligned}$$

IPCW-(Iterative) Targeted MLE based on reduced data at specified g^r: We will now compute the iterative targeted MLE under i.i.d sampling $O_1^r, \ldots, O_n^r$ from $P_{Q_0^r,g^r}^r$, treating g^r as known (e.g., estimated a priori), but assigning IPCW-weights, as follows. Let Q^{r0} be an initial estimator of Q_0^r such as a weighted-MLE according to a working model $\mathcal{Q}_j^r$:

$$Q_j^{r0} = \arg\max_{Q_j^r \in \mathcal{Q}_j^r} \sum_i \log Q_j^r(O_i^r) w_{ji}.$$

Compute the overall amount of fluctuation with weighted maximum likelihood estimation:

$$\epsilon_n^1 = \arg\max_{\epsilon} \sum_i \sum_j \log Q_j^{r0}(\epsilon)(O_i^r) w_{ji},$$

and compute the corresponding first step targeted ML update $Q_j^{r1} = Q_j^{r0}(\epsilon_n^1)$, $j = 1, \ldots, J$, and thereby the overall update $Q^{r1} = Q^{r0}(\epsilon_n^1)$. Iterate this process till convergence (i.e., $\epsilon_n^k \approx 0$) and denote the final update with $Q_n^r = (Q_{jn}^r : j = 1, \ldots, J)$.

Let $D(Q^r, g^r, g_0) = \sum_j D_j^r(Q^r, g^r)\frac{g_j^r}{g_{0j}}$. Under a weak regularity condition we have (see proof in **?**)

$$0 = \sum_i D(Q_n^r, g^r, g_0)(O_i) = \sum_i \sum_j D_j^r(Q_n^r, g^r)(O_i^r) w_{ji}. \tag{7}$$

Substitution estimator: Our estimator of ψ_0 is given by $\Psi^r(Q_n^r)$.

The IPCW-R-TMLE is an estimator Q_n^r solving an IPCW-reduced data efficient influence curve equation (7). Firstly, we establish that this IPCW-reduced data efficient influence curve is an "estimating function" for the target

parameter with nice robustness properties w.r.t its nuisance parameters Q_0^r and g_0. Subsequently, we discuss the corresponding implications on the statistical properties of the IPCW-R-TMLE.

Robustness properties of IPCW-Reduced Data Efficient Influence Function: Recall that $D^r(Q^r, g^r)$ denotes the efficient influence curve for the reduced data $O^r \sim P_{Q^r,g^r}$ for model $\mathcal{M}^r$ and parameter Ψ^r. It follows from general results in **?** that $P_{Q_0^r,g_0^r} D^r(Q^r, g^r) = 0$ if either $Q^r = Q_0^r$ or $\Psi(Q^r) = \Psi(Q_0^r)$ and $g^r = g_0^r$. This double robustness result for D^r is exploited/inherited by the estimating function

$$D(Q^r, g^r, g_0) \equiv \sum_j D_j^r(Q^r, g^r) g_j^r / g_{0j},$$

whose corresponding estimating equation is solved by our IPCW targeted MLE, in the following manner. We have

$$\begin{aligned} P_{Q_0,g_0} D(Q^r, g^r, g_0) &= P_{Q_0,g_0} \sum_j D_j^r(Q^r, g^r) \frac{g_j^r}{g_j} \\ &= P_{Q_0,g^r} \sum_j D_j^r(Q^r, g^r) \frac{g_{0j}}{g_j}. \end{aligned}$$

This implies that if $g_j = g_{0j}$ (i.e., the action mechanism is correctly specified), then $P_{Q_0,g_0} D(Q^r, g^r, g_0) = 0$ for all choices of Q^r, g^r with $\Psi(Q^r) = \Psi(Q_0^r)$. In a typical scenario, we have that Q_{j0}^r denotes the conditional distribution of $L^r(j^r)$, given $A(0), \ldots, A(j^r - 1)$ and $\bar{L}^r(j^r - 1)$. In this case, if g_{0j} is only a function of O^r, then if $Q^r = Q_0^r$, it follows that $P_{Q_0,g^r} D_j^r(Q_0^r, g^r) \frac{g_{0j}}{g_j} = 0$ for all g_j only being a function of O^r (by using that the conditional expectation of a score $D_j^r(Q_0^r, g^r)$ of Q_{j0}^r, given $(A(0), \ldots, A(j^r - 1)$ and $\bar{L}^r(j^r - 1)$, equals zero), and as a consequence, $P_{Q_0,g_0} D(Q_0^r, g^r, g) = 0$ for such misspecified g. That is, in the case that the true g_0 and its asymptotic fit are only functions of the reduced data structure, we have the double robustness of the estimating function $D(Q^r, g^r, g)$ in the sense that $P_{Q_0,g_0} D(Q^r, g^r, g) = 0$ if either $Q^r = Q_0^r$ or $g = g_0$, for all g^r.

Statistical Properties of IPCW-Reduced Data Targeted MLE: The above mentioned robustness property of the estimating equation $\sum_i D(Q_n^r, g_n^r, g_n) = 0$, g_n an estimator of g_0, as solved by the IPCW-R-TMLE Q_n^r translates under regularity conditions in the following statistical properties of the substitution estimator $\psi_n = \Psi^r(Q_n^r)$. Firstly, under appropriate regularity conditions, if g_n consistently estimates g_0, then ψ_n will be a consistent and asymptotically linear estimator of ψ_0. In addition, if $g_n(A \mid X)$ and its

target $g_0(A \mid X)$ are only functions of the reduced data structure O^r, then 1) ψ_n is consistent and asymptotically linear if either Q_n^r consistently estimates Q_0^r or g_n consistently estimates g_0, and if both estimates are consistent, then the estimator ψ_n is more efficient than an efficient estimator based on n i.i.d. observations of the reduced data structure O^r only.

21 IPCW-Reduced Data-Targeted-MLE for Marginal Structural Models.

Let $O = (W = L(0), A(0), \ldots, L(K), A(K), Y = L(K+1))$, $L(0)$ are baseline co-variates, $A(j) = (A_1(j), A_2(j))$, $A_1(j)$ denotes a treatment at time j, $A_2(j) = I(C \leq j)$ indicates a censoring event/drop out at time j, $L(j)$ are time dependent co-variates collected after $A(j-1)$ and before $A(j)$, and Y is a final outcome of interest collected at time $K+1$. The chronological ordering of the data implies that $L(j) = L_{\bar{A}(j-1)}(j)$ is affected by past action history $\bar{A}(j-1)$. Let the full data structure be $X = (L_a : a \in \mathcal{A})$, $L_a(t) = L_{\bar{a}(t-1)}(t)$, so that the observed data structure O can be presented as a missing data structure $O = (A, L_A)$. We assume the sequential randomization assumption $g_0(A(j) \mid \bar{A}(j-1), X) = g(A(j) \mid \bar{A}(j-1), \bar{L}(j))$, $j = 0, \ldots, K$. We have $O \sim dP_{Q_0,g_0}(A, L) = Q_0(A, L)g_0(A \mid X)$, where $Q_0(a, l) = P(L_a = l)$, under the assumption that $g(a \mid X) > 0$ for all $a \in \mathcal{A}$.

Consider a marginal structural working model $E_0(Y_{a_10} \mid V) = m(a_1, V \mid \beta_0)$ for a user supplied working model $\{m(\cdot \mid \beta) : \beta\}$ for the counterfactual mean of Y_{a_10} under treatment regimen $a_1 = (a_1(0), \ldots, a_1(K))$ and no censoring (i.e., $a_2 = 0$), conditional on baseline covariates V included in the set of baseline covariates $W = L(0)$. Our goal is to estimate β_0 defined non-parametrically as

$$\beta_0 = \Psi(Q_0) \equiv \arg\min_{\beta} E_{Q_0} \sum_{a_1} h(a_1, V)(m(a_1, V \mid \beta) - E_{Q_0}(Y_{a_10} \mid V))^2$$

for some user supplied weight function $h(a_1, V)$. A typical choice is $h(a_1, V) = g^*(a_1 \mid V)$, where g^* is a conditional distribution of A_1, given V, representing the limit of an estimate of the true conditional distribution of A_1, given V according to a possibly misspecified working model. Equivalently,

$$\beta_0 = \Psi(Q_0) = \arg\min_{\beta} E_{Q_0} \sum_{a_1} h(a_1, V)(Q_0(a_1, W) - m(a_1, V \mid \beta))^2,$$

where we define $Q_0(a_1, W) = E_0(Y_{a_10} \mid W)$.

The model for the observed data structure $O \sim dP_{Q_0,g_0} = Q_0 g_0$ can be written as $\mathcal{M} = \{P_{Q,g} : Q, g \in \mathcal{G}\}$, where Q can be arbitrary and $\mathcal{G}$ is the set of conditional distributions of A, given X, satisfying SRA.

Data reduction: Let the reduced data be obtained by excluding all the time-dependent co-variates $O^r = (W, A = (A(0), \ldots, A(K)), Y_A)$. Let $X^r = (W, (Y_a : a \in \mathcal{A}))$, so that $O^r = (W, A, Y_A)$ is a missing data structure on X^r.

SRA for reduced data: Consider an action mechanism g^r satisfying $g^r(A \mid X) = g^r(A \mid X^r) = g^r(A \mid W)$. We consider a choice g^r so that $P(A_2 = 0) = 1$ under g^r.

Reduced Data Model: In the reduced data model for O^r one assumes $g^r(A \mid X^r) = g^r(A \mid W)$, so that $O^r \sim p_{Q_0^r,g^r} = Q_0^r g^r$, $Q_0^r = Q_{01}^r * Q_{02}^r$, where Q_{01}^r is a marginal distribution of W, Q_{02}^r is a conditional distribution of Y, given A, W, and g^r is the conditional distribution of A, given X^r. We have $Q_{02}^r(y \mid a, w) = P(Y_a = y \mid W = w)$. Let $\mathcal{M}^r = \{p_{Q^r,g^r} : Q^r, g^r \in \mathcal{G}^r\}$, where $\mathcal{G}^r = \{g(\cdot \mid X^r) = g(\cdot \mid W)\}$ is the class of conditional distributions of A, given X^r, only depending on X^r through W. We note that Q_0^r is a function of Q_0, and both are identified as counterfactual distributions: $Q_0^r(w, a, y) = P(W = w, Y_a = y)$ is a sub-distribution of $Q_0(a, l) = P(L_a = l)$.

Consider the parameter

$$\beta_0^r = \Psi^r(Q_0^r) \equiv \arg\min_{\beta} E_0 \sum_{a_1} h(a_1, V)\{Q_0^r(a_1, W) - m(a_1, V \mid \beta)\}^2,$$

where $Q_0^r(a_1, W) = E_0(Y_{a_1 0} \mid W)$. It follows that $\beta_0^r = \beta_0$. In general, $\Psi^r(Q^r) = \Psi(Q)$ for any Q and corresponding Q^r.

Efficient influence curve of Ψ^r in reduced data model: The efficient influence curve of Ψ^r at $p_{Q_0^r,g^r} \in \mathcal{M}^r$ is given by

$$\begin{aligned} D^r &= \frac{h(A_1, V)}{g^r(A_1 0 \mid W)} \frac{d}{d\beta_0} m(A_1, V \mid \beta_0)(Y - Q_0^r(A_1, W)) I(A_2 = 0) \\ &\quad + \sum_{a_1} h(a_1, V) \frac{d}{d\beta_0} m(a_1, V \mid \beta_0)(Q_0^r(a_1, W) - m(a_1, V \mid \beta_0)) \\ &\equiv D_1^r(Q_0^r, g^r) + D_2(Q_0^r), \end{aligned}$$

where $Q_0^r(a_1, W) = E_0(Y_{a_1 0} \mid W)$. We also note that g^r can be factored as

$$\begin{aligned} g^r(A_1 0 \mid W) &= \prod_{j=0}^{K} g_1^r(A_1(j) \mid A_2(j) = 0, \bar{A}_1(j), W) \\ &\quad \prod_{j=1}^{K} g_2^r(A_2(j) = 0 \mid \bar{A}_1(j-1), A_2(j-1) = 0, W), \end{aligned}$$

where g_1^r represents a treatment mechanism and g_2^r a censoring mechanism.

IPCW-Weighted Reduced Data Efficient Influence Curve: By weighting the first component D_1^r with g^r/g_0, we obtain

$$D_1(Q_0^r, g^r)g^r(A \mid X^r)/g_0(A \mid X) = D_1(Q_0^r, g^r)g^r(A_1 0 \mid X^r)/g_0(A_1 0 \mid X),$$

which yields the following IPCW-Weighted Reduced Data Efficient Influence Curve:

$$\begin{aligned} D(Q_0, g_0) &= \frac{h(A_1, V)}{g_0(A_1 0 \mid X)} \frac{d}{d\beta_0} m(A_1, V \mid \beta_0)(Y - Q_0^r(A_1, W))I(A_2 = 0) \\ &\quad + \sum_{a_1} h(a_1, V) \frac{d}{d\beta_0} m(a_1, V \mid \beta_0)(Q_0^r(a_1, W) - m(a_1, V \mid \beta_0)) \\ &\equiv D_1(Q_0^r, g_0) + D_2(Q_0^r). \end{aligned}$$

IPCW-R-Targeted MLE solving IPCW-Reduced Data Efficient Influence Curve Equation: We will now compute the iterative targeted MLE under i.i.d sampling $O_1^r, \ldots, O_n^r$ from $P^r_{Q_0^r, g^r}$, treating g^r as known, but assigning IPCW-weights, as follows. Firstly, we estimate the marginal distribution Q_{01}^r of W with the empirical probability distribution of $W_1, \ldots, W_n$. Let Q_2^{r0} be an initial estimator of the conditional distribution Q_{20}^r of $Y_{a_1 0}$, given W, such as a weighted-MLE according to a working model for Y, given A, W:

$$Q_2^{r0} = \arg \max_{Q_2^r \in \mathcal{Q}_2^r} \sum_i \log Q_2^r(Y_i \mid A_i, W_i) w_i,$$

where

$$w_i = I(A_{2i} = 0) \frac{g_n^r(A_{1i} 0 \mid W_i)}{g_n(A_{1i} 0 \mid X_i)}.$$

For example, if one assumes a normal error regression model for Y on A, W, then this corresponds with weighted least squares regression, and if Y is binary, and one assumes a logistic regression model for Y, given A, W, then this corresponds with weighted logistic linear regression.

Subsequently, we extend the current fit Q_2^{r0} with an ϵ-extension so that the score at $\epsilon = 0$ equals $D_1(Q^{r0}, g^r)$. As shown previously, in the normal regression model case, this corresponds with adding a covariate-extension

$$\epsilon \frac{h(A_{1i}, V_i)}{g^r(A_{1i} 0 \mid W_i)} \frac{d}{d\beta_0} m(A_{1i}, V_i \mid \beta_0)$$

and, in the logistic regression case, one adds the covariate extension

$$\epsilon \frac{h(A_{1i}, V_i)}{g^r(A_{1i} 0 \mid W_i)} \frac{\frac{d}{d\beta_0} m(A_{1i}, V_i \mid \beta_0)}{m_{\beta_0}(1 - m_{\beta_0})(A_{1i}, V_i)}$$

to the logit. We now compute the amount of fluctuation with weighted maximum likelihood

$$\epsilon_n^1 = \arg\max_{\epsilon} \sum_i \sum_j \log Q_2^{r0}(\epsilon)(O_i^r) w_i,$$

which corresponds with univariate weighted least squares regression or univariate weighted logistic regression, and can thus be done with standard software.

We now compute the corresponding first step targeted ML update $Q_2^{r1} = Q_2^{r0}(\epsilon_n^1)$. We iterate this process till convergence (i.e., $\epsilon_n^k \approx 0$) and denote the final update with Q_{2n}^r. If $m(\cdot \mid \beta)$ is linear in β or if it is a logistic linear model, then it follows that the ϵ-extensions mentioned above do not depend on the updates Q_2^{rk}, and, as a consequence, convergence occurs in one single update step: $Q_2^{rk} = Q_2^{r1}$, $k = 2, 3, \ldots$.

Let $Q_n^r = (Q_{1n}^r, Q_{2n}^r)$ be the corresponding estimate of the true $Q_0^r = (Q_{01}^r, Q_{02}^r)$. Under a weak regularity condition we have that the IPCW-R-TMLE Q_n^r of Q_0^r solves the IPCW-R-Efficient influence curve equation

$$0 = \sum_i D(Q_n^r, g_n)(O_i).$$

Substitution estimator: The IPCW-R-TML estimator of $\beta_0 = \Psi(Q_0)$ is given by $\Psi^r(Q_n^r)$.

Estimation of Treatment and Censoring mechanism: When estimating $g_0(A \mid X)$ it is a good strategy to give preference to the baseline covariates W, so that the time-dependent covariates are only entered if they provide significant improvement relative to a fit based on the baseline covariates only. In this manner, one obtains relatively stable weights $w_i = g_0^r(A_i \mid X_i)/g_0(A_i \mid X_i)$. In addition, as point out above, it exploits maximally the double robust property of the IPCW-R-Efficient influence curve function w.r.t. the baseline covariates.

22 IPCW-Reduced Data-Targeted-MLE for Marginal Structural hazard models

Let $O = (W = L(0), A(0), \ldots, L(K), A(K), L(K+1))$, $L(0)$ are baseline covariates, $A(j) = (A_1(j), A_2(j))$, $A_1(j)$ denotes a treatment at time j, $A_2(j) = I(C \leq j)$ indicates a censoring event/drop out at time j, $L(j)$ are time dependent co-variates collected after $A(j-1)$ and before $A(j)$, and $L(j)$ includes a survival component $Y(j) = I(T \leq j)$. The observed data structure becomes degenerate after a censoring or survival event.

We have that $L(j) = L_{\bar{A}(j-1)}(j)$ is affected by past action history $\bar{A}(j-1)$. Let the full data structure be $X = (L_a : a \in \mathcal{A})$. We have that $L_a(t) = L_{\bar{a}(t-1)}(t) = L_{\bar{a}(t-1)}(\min(t, T_a))$ is truncated at the survival time, and $L_a(t)$ includes the survival component $Y_a(t) = I(T_a \leq t)$ itself. The observed data structure O can be presented as a missing data structure $O = (A, L_A)$. We assume the sequential randomization assumption $g_0(A(j) \mid \bar{A}(j-1), X) = g(A(j) \mid \bar{A}(j-1), \bar{L}(j))$, $j = 0, \ldots, K$. We have $O \sim dP_{Q_0,g_0}(A, L) = Q_0(A, L)g_0(A \mid X)$, where $Q_0(a, l) = P(L_a = l)$, under the assumption that $g(a \mid X) > 0$ for all $a \in \mathcal{A}$.

Consider a marginal structural model $E_0(dY_{a_10}(t) \mid \bar{Y}_{a_10}(t-), V) = Y^*_{a_10}(t)\lambda_{\beta_0}(t, a_1, V)$ for a user supplied working model $\{\lambda_\beta : \beta\}$ for the intensity of Y_{a_10} under treatment regimen $a_1 = (a_1(0), \ldots, a_1(K))$ and no censoring, conditional on baseline covariates V included in the set of baseline covariates $W = L(0)$. The model for the observed data structure $O \sim dP_{Q_0,g_0} = Q_0g_0$ can be written as $\mathcal{M} = \{P_{Q,g} : Q, g \in \mathcal{G}\}$, where Q can be arbitrary and $\mathcal{G}$ is the set of conditional distributions of A, given X, satisfying SRA.

Data reduction: Let the reduced data $O^r = (W, A = (A(0), \ldots, A(K)), (Y_A(t) : t))$ be obtained by excluding all the time-dependent co-variates . Let $X^r = (W, (Y_a : a \in \mathcal{A}))$, so that $O^r = (W, A, Y_A)$ is a missing data structure on X^r.

SRA-Reduced Data Model: In the reduced data model for O^r one assumes that the conditional probability distribution $g^r(A \mid X^r)$ of A, given X^r, satisfies the SRA assumption w.r.t. X^r (i.e., $g^r(A \mid X^r)$ is a measurable function of W, A, Y), so that $O^r \sim p_{Q^r_0,g^r} = Q^r_0 g^r$, $Q^r_0 = Q^r_{01} * \prod_j Q^r_{02j}$, where Q^r_{01} is a marginal distribution of W, Q^r_{02j} is a conditional distribution of $Y(j)$, given $\bar{Y}(j-1), \bar{A}(j-1), W$, and g^r is the conditional distribution of A, given X^r. Let $\mathcal{M}^r = \{p_{Q^r,g^r} : Q^r, g^r \in \mathcal{G}^r\}$, where $\mathcal{G}^r$ is the class of conditional distributions of A, given X^r satisfying the SRA w.r.t X^r. We note that Q^r_0 is a function of Q_0, and both are identified as counterfactual distributions: Q^r_0 identifies the distribution of (T_a, W) and is thus a sub-distribution of Q_0, since Q_0 identifies the marginal distribution of L_a.

Below, we first present the targeted MLE based on the reduced data structure under i.i.d. sampling from $P^r_{Q^r_0,g^r}$, and subsequently we show how to bring in the IPCW-weights to obtain the IPCW-R-TMLE.

Reduced data Targeted MLE for a marginal structural logistic regression model for survival outcome:

For each treatment strategy $a_1 \in \mathcal{A}$, let T_{a_10} be a treatment specific counterfactual survival time, and let the full data on each experimental unit be given by

$(W, (T_{a_1 a_2} : a \in \mathcal{A}))$. Suppose we observe $O = (W, A, T = T_A)$. Suppose that the survival times are discrete on time points indexed by $j = 0, 1, \ldots$. Consider the following class of causal working models for the treatment specific hazard:

$$P(T_{a_1 0} = t \mid T_{a_1 0} \geq t, V) = \lambda_{\beta_0}(a_1, t, V),$$

for a given working model $\lambda_\beta(a_1, t, V)$ indexed by parameter vector β. Let $dY(t) = I(T = t)$ and $dY_{a_1 0}(t) = I(T_{a_1 0} = t)$. The typical working model will be a logistic regression model:

$$\lambda_\beta(a_1, t, V) = \frac{1}{1 + \exp(-m_\beta(t, a_1, V))},$$

where $m_\beta(t, a_1, V)$ is a specified function linear in summary measures of $(\bar{a}_1(t-1), t, V)$. We also assume that A_1 is independent of $X = (W, (Y_{a_1 0} : a_1))$, given W, and $P(A_2 = 0) = 1$.

The class of so called IPTW-estimating functions for β_0 are given by:

$$D_{IPTW,h} = \sum_t I(C > t) \frac{h(A_1, t, V)}{g^r(\bar{A}_1(t-1), C > t \mid X)} \times \frac{d}{d\beta_0} \lambda_{\beta_0}(t, A_1, V)(dY(t) - I(T \geq t)\lambda_{\beta_0}(t, A_1, V)). \quad (8)$$

By projecting the $D_{IPTW,h}$ on the tangent space of the relevant (i.e., ignoring the treatment mechanism) factor of the likelihood of O, given by

$$P(W) \prod_t P(dY(t) \mid \bar{Y}(t-1), \bar{A}(t-1), W),$$

we obtain the efficient influence curve of $\beta_0 = \beta_{0h}$ defined non-parametrically as the solution of $P_0 D_{IPTW,h}(\beta, g_0) = 0$.

This yields the following representation of this efficient influence curve

$$\begin{aligned} D_h^*(\beta_0, Q_0^r, g^r) &= \sum_{t=0} h^*(Q_0^r, g^r)(t, \bar{A}(t-1), W)(dY(t) \\ &\quad - E(dY(t) \mid \bar{Y}(t-1), \bar{A}(t-1), W)) + E(D_{IPTW,h} \mid W) \\ &\equiv D_1(Q_0^r, g_0^r)(W, A, Y) + D_2(Q_0^r)(W), \end{aligned}$$

where

$$\begin{aligned} h^* &= \{E_{Q_0^r, g_0^r}(D_{IPTW,h} \mid dY(t) = 1, \bar{Y}(t-1), \bar{A}(t-1), W) \\ &\quad - E_{Q_0^r, g_0^r}(D_{IPTW,h} \mid dY(t) = 0, \bar{Y}(t-1), \bar{A}(t-1), W)\}, \end{aligned}$$

where D_2 represents a score of the marginal distribution of W and the first term D_1 represents a sum over t of scores of $P(dY(t) \mid \bar{Y}(t-1), \bar{A}(t-1), W)$. In the special case that $V = W$ and an appropriate choice of h we have that $D_h^* = D_{IPTW,h}$ since $D_{IPTW,h}$ is already an element of the tangent space.

The second term defines $\beta(Q)$ as a function of Q through the following least squares solution (check):

$$\beta(Q) = \arg\min_{\beta} E_Q \sum_a \sum_t h(t, \bar{a}_1(t-1), V) \left\{ q(t, a_1, V) - \bar{Q}(t, a_1, V) \lambda_\beta(t, a_1, V) \right\}^2,$$

and for Q_{1n} being the empirical distribution of $W_1, \ldots, W_n$ this gives us:

$$\beta(Q_{1n}, Q_{2n}) = \arg\min_{\beta} \sum_i \sum_a \sum_t h(a_1, t, V_i) \left\{ q_n(t, a_1, V_i) - \bar{Q}_n(t, a_1, V_i) \lambda_\beta(t, a_1, V_i) \right\}^2.$$

Here $q(t, a_1, V) = E(dY_{a_1 0}(t) \mid V)$ and $\bar{Q}(t, a_1, V) = E(I(T_{a_1 0} \geq t) \mid V)$. In other words, the choice of h defines β_0 as a weighted projection of of the true hazard $q_0/\bar{Q}_0$ on the working model $\{\lambda_\beta(a_1, t, V) : \beta\}$.

Consider an initial fit of λ^0 of $E(dY(t) \mid \bar{Y}(t-1) = 0, \bar{A}(t-1), W)$ based on a logistic regression model:

$$\lambda^0 = I(T_A \geq t) \frac{1}{1 + \exp(-m^0)}.$$

Consider the following ϵ-extension:

$$\lambda^0(\epsilon) = \frac{1}{1 + \exp(-m^0 - \epsilon h^*(Q^{0r}, g^r))}.$$

The score of $\lambda^0(\epsilon)$ at $\epsilon = 0$ equals the wished component $D_2(\beta(Q^{r0}), Q^{r0}, g^r)$ of the efficient influence curve. Thus, assuming an initial fit Q^0 for which Q_1^0 is the empirical distribution of $W_1, \ldots, W_n$, it follows that the with $\lambda^0(\epsilon)$ corresponding $Q^0(\epsilon)$ (and no update of the already nonparametric MLE Q_1^0) has score at $\epsilon = 0$ equal to the efficient influence curve $D^*(\beta^0, Q^{r0}, g^r)$.

The iterative targeted-MLE: This defines the wished ϵ-extension $Q^0(\epsilon)$ of an initial fit Q^0. Let ϵ_n^0 be the MLE over ϵ for $Q^0(\epsilon)$. Let $Q_n^1 = Q_n^0(\epsilon_n^0)$ be the updated estimate which corresponds with an updated $\beta_n^1 = \beta_h(Q_n^1)$. We iterate this updating process till the corresponding sequence β_n^k is such that $\beta_n^k - \beta_n^{k-1}$ does not significantly change anymore.

We denote the selected final update with $Q_n = Q_n^{k*}$ for some k^*, and $\beta_n = \beta_n^{k*}$, respectively, and we refer to this estimate β_n as the (iterative) targeted MLE of β_0.

Important special case: $W = V$. In this special case and by setting h so that it cancels out $g^r(\bar{A}_1 \mid X^r)$, it follows that the targeted MLE is nothing else than the MLE which corresponds with fitting a logistic regression of $E(dY(t) \mid \bar{Y}(t-1), \bar{A}(t-1), V) = I(T \geq t) \lambda_\beta(t, A_1, V)$.

IPCW-Reduced data targeted MLE.

We have the following IPCW-reduced data efficient influence curve

$$\begin{aligned}
&D_h^*(\beta_0, Q_0^r, g^r, g_0) = \\
&\textstyle\sum_{t=0} I(C>t)\frac{g^r(\bar{A}_1(t-1),C>t|X^r)}{g_0(\bar{A}_1(t-1),C>t|X)} h^*(Q_0^r, g^r)(t, \bar{A}_1(t-1), W) \\
&\times (dY(t) - E(dY(t) \mid \bar{Y}(t-1), \bar{A}_1(t-1), W)) + E(D_{IPTW} \mid W) \\
&\equiv D_1(Q_0^r, g_0^r)(W, A, Y) + D_2(Q_0^r)(W),
\end{aligned}$$

where

$$\begin{aligned}
h^* \;=\; & E_{Q_0^r, g_0^r}(D_{IPTW} \mid dY(t) = 1, \bar{Y}(t-1), \bar{A}(t-1), W) \\
& - E_{Q_0^r, g_0^r}(D_{IPTW} \mid dY(t) = 0, \bar{Y}(t-1), \bar{A}(t-1), W).
\end{aligned}$$

The corresponding implementation of the IPCW-reduced data targeted MLE corresponds with computing the reduced data targeted MLE under the assumption that A, given X^r, follows distribution g^r, but where we used IPCW weights for each time point $I(C > t)g^r(\bar{A}_1(t-1), C > t \mid X^r)/g_0(\bar{A}_1(t-1), C > t \mid X)$. A particular attractive approach is to set $W = V$ (i.e., let the reduced data structure only contain baseline covariates V), since in that case the reduced data targeted MLE is just a regular MLE according to the working model for the hazard, so that the IPCW-reduced data T-MLE is obtained as an equally easy to compute IPCW-weighted MLE. That is, one fits a standard logistic regression hazard model based on the baseline covariates V using the time-dependent IPCW-weights, and subsequently one evaluates the obtained fit to obtain the wished estimator of β_0.

23 Targeted Empirical Bayesian Learning.

The iterative targeted maximum likelihood estimation methodology resulting in a sequence of updated density estimators converging to a solution of the efficient influence curve equation can be generalized to a targeted empirical Bayesian learning method in which one assumes a prior distribution on the parameter of interest and ends up with a targeted posterior distribution of this parameter of interest.

Consider the setting in which we observe n i.i.d. observations $O_1, \ldots, O_n$ of a random variable $O \sim P_0$, which is known to be an element of a model $\mathcal{M}$, and let $\Psi : \mathcal{M} \to \mathbb{R}^d$ be the target parameter mapping of interest.

Step 1, Determine Prior Distribution on Parameter of Interest: Specify a prior distribution Π of the parameter ψ_0. Let f_Π be the density of Π.

Step 2, Determine targeted (frequentist) estimator of distribution P_0: Consider an (e.g., initial) estimated probability distribution $\hat{P}$ in the model $\mathcal{M}$. This estimator is recommended to be a targeted estimator itself such as the iterative targeted MLE.

Step 3, Determine targeted fluctuation function: Let $\{\hat{P}(\epsilon) : \epsilon\} \subset \mathcal{M}$ be a fluctuation through $\hat{P}$ at $\epsilon = 0$ with score at $\epsilon = 0$ equal to the efficient influence curve $D^*(\hat{P})$ at $\epsilon = 0$.

Step 4, Derive prior distribution on ϵ equivalent with prior on ψ_0: Determine a prior distribution on ϵ that yields the assumed prior distribution on the true parameter ψ_0 of interest. For this purpose one notes that a prior distribution on a set $\mathcal{E}$ of ϵ-values implies a prior distribution on $\{\Psi(\hat{P}(\epsilon)) : \epsilon \in E\}$ (and thus on ψ_0) through the mapping $f(\hat{P}) : \epsilon \rightarrow \Psi(\hat{P}(\epsilon))$. As a consequence, one can choose the prior distribution of ϵ as the probability distribution of $f(\hat{P})^{-1}(X)$ with $X \sim \Pi$, assuming $f(\hat{P})$ is invertible. This corresponds with a random variable E defined by drawing from Π and applying $f(\hat{P})^{-1}$ to it. Let Π^* be this prior distribution of ϵ. The density of Π^* is given by

$$f_{\Pi^*}(\epsilon) = f_{\Pi}(f(\hat{P})(\epsilon))J(\epsilon),$$

where $J(\epsilon) =\mid \frac{d}{d\epsilon} f(\hat{P})(\epsilon) \mid$ is the Jacobian corresponding with transformation $\psi = f(\hat{P})(\epsilon)$.

Step 5, Determine (targeted) posterior distribution of ϵ, given data, treating $\hat{P}$ as fixed/non random: Since, from a Bayesian perspective, the conditional density of $O_1, \ldots, O_n$, given ϵ, is given by $\prod_{i=1}^n d\hat{P}(\epsilon)(O_i)$, by Bayes formula, the posterior density of ϵ, given the data $O_1, \ldots, O_n$, treating $\hat{P}$ as fixed and given, is given by (up till normalizing constant)

$$\propto \prod_{i=1}^n d\hat{P}(\epsilon)(O_i) f_{\Pi^*}(\epsilon).$$

One can use standard Bayesian methodology such as Monte-Carlo Markov Chain sampling to sample a large number of draws, say, $E_1, \ldots, E_B$, from this posterior distribution of ϵ, given $O_1, \ldots, O_n$.

Step 6, Output targeted posterior distribution of ψ_0, given data, treating $\hat{P}$ as fixed/non random: The posterior distribution of ψ_0 is now described by the sample $f(\hat{P})(E_b) = \Psi(\hat{P}(E_b))$, $b = 1, \ldots, B$.

Optional: Iterate. If $\hat{P}$ was not a targeted estimator, then one could compute the posterior mean of ϵ, given $O_1, \ldots, O_n$, and compute the updated distribution $P^1 = \hat{P}(E^0(\epsilon \mid O_1, \ldots, O_n))$ by substituting the posterior mean of ϵ into the fluctuation function $\hat{P}(\epsilon)$ for ϵ. One now carries out Step 3-5 (thus with the same a priori specified prior distribution on ψ_0) and one iterates this process till the posterior mean of ϵ converges to zero at which point we have achieved out wished targeted estimator of P_0. One now finalizes the procedure with Step 6, by outputting the posterior distribution of ψ_0.

We refer to this methodology as *empirical* targeted Bayesian because we treat the (initial) frequentist estimator $\hat{P}$ in the model $\{\hat{P}(\epsilon) : \epsilon\}$ for the data generating distribution as fixed so that only ϵ is treated as a parameter on which we put a prior distribution, and we calculate its posterior distribution accordingly.

Rational behind the targeted posterior distribution on parameter of interest: The rational of this methodology for generating a posterior distribution of ψ_0 is as follows. To evaluate the posterior distribution of ψ_0 we need to be concerned about its bias w.r.t to ψ_0 and its spread needs to be representative of the actual standard error of the posterior mean. Regarding the bias, because $\hat{P}$ is a targeted estimator of the data generating distribution such as the iterative targeted MLE, $\Psi(\hat{P})$ is a robust and locally efficient estimator of ψ_0. Consequently, also the posterior mean of the outputted posterior distribution of ψ_0 will be centered closely around $\Psi(\hat{P})$ and will thus represent a robust and locally efficient estimator w.r.t to frequentist theory. Regarding the spread, one needs to know that $\{P_0(\epsilon) : \epsilon\}$, whose score at $\epsilon_0 = 0$ equals the efficient influence curve of $\Psi : \mathcal{M} \to \mathbb{R}^d$ at P_0, is a so called hardest submodel for estimation of ψ_0 (e.g., see **?** or **?**) in the sense that estimation of the parameter $\psi_0 = \Psi(P_0(\epsilon_0))$ of ϵ_0 in this hardest sub-model of $\mathcal{M}$ is asymptotically as hard as it is to estimate ψ_0 in the actual model $\mathcal{M}$. As a consequence, statistical inference (i.e., asymptotic covariance matrix, and information matrix) for a maximum likelihood estimator of $\psi_0 = \Psi(P_0(\epsilon_0))$ in this sub-model will be representative of the actual estimator $\Psi(P^0(E(\epsilon \mid O_1, \ldots, O_n))$ of ψ_0.

23.1 Example: Targeted Bayesian learning of Survival function.

We now illustrate this completely general empirical targeted Bayesian analogue of the iterative targeted MLE methodology with a simple example. Suppose,

we wish to estimate a survival function at a point, $\psi_0 = P_0(O > x_0)$, based on n i.i.d. observations $O_1, \ldots, O_n \sim P_0$ in a nonparametric model for P_0.

Prior distribution: Consider a prior distribution on ψ_0 such as a uniform distribution on $[a_0, b_0] \subset [0, 1]$ for some numbers $0 \leq a_0 < b_0 \leq 1$. Let π be its density.

Targeted density estimator: Consider a targeted ML density estimator $\hat{p} = p^0(\epsilon^0)$ of the density p_0, given an initial density estimator p^0, a one-dimensional fluctuation function $\epsilon \to p^0(\epsilon)$ (into valid densities), and

$$\epsilon^0 = \arg\max_{\epsilon} \prod_{i=1}^{n} p^0(\epsilon)(O_i),$$

satisfying $0 = \sum_i D^*(p^0(\epsilon^0))(O_i)$, where $D^*(p) = I(O \leq x_0) - \int_{x_0}^{\infty} p(x)dx$ is the efficient influence curve of Ψ at p. In **?** we showed that indeed the first step targeted MLE's can be constructed to already solve the efficient influence curve estimating equation: e.g. choose $p^0(\epsilon) = (1 + \epsilon D^*(p^0)(O))p^0$.

Targeted fluctuation of targeted density estimator: Let $\hat{p}(\epsilon) = (1 + \epsilon D^*(\hat{p}))\hat{p}$ be the targeted fluctuation function of $\hat{p}$ whose score at $\epsilon = 0$ indeed equals $D^*(\hat{p})$.

Evaluate prior distribution for ϵ implied by prior of ψ_0: We have $\epsilon \to f(\hat{P})(\epsilon) = \Psi(\hat{p}(\epsilon)) = \int_{x_0}^{\infty}(1 + \epsilon D^*(\hat{P})(x))\hat{p}(x)dx$. The inverse of $\epsilon \to f(\hat{P})(\epsilon)$ is given by

$$g(\psi) \equiv f(\hat{P})^{-1}(\psi) = \frac{\psi - \Psi(\hat{p})}{E_{\hat{p}} D^{*2}(\hat{p})},$$

which shows that $f(\hat{P})$ is invertible. In particular, this shows that we can choose the prior distribution of ϵ as the distribution of $g(X)$ with $X \sim \Pi$, where Π is the prior distribution on ψ_0 specified initially.

Targeted posterior density of ϵ: The derivative of $f(\hat{P})$ at ϵ is given by $\sigma^2 \equiv E_{\hat{p}} D^{*2}(\hat{P})$ so that the Jacobian is given by a constant $J(\epsilon) = \sigma^2$. The univariate posterior density of ϵ, given $O_1, \ldots, O_n$, is thus given by

$$\pi(\epsilon \mid O_1, \ldots, O_n) = \frac{\prod_{i=1}^{n} \hat{p}(\epsilon)(O_i)\pi(f(\hat{P}(\epsilon))}{\int_{\epsilon} \prod_{i=1}^{n} \hat{p}(\epsilon)(O_i)\pi(f(\hat{P}(\epsilon))}, \tag{9}$$

where we recall that π is the density of the prior distribution on ψ_0.

Targeted posterior density of survival function: The posterior density of ϵ implies the posterior distribution of $f(\hat{P}(\epsilon)) = \Psi(\hat{P}(\epsilon))$, i.e. the survival function at x_0. In this example, one can even pursue analytic calculation of this posterior density of the survival function since it only involves

univariate density calculations. The Monte Carlo simulation approach would be to sample $E_1, \ldots, E_B$ from the posterior density $\pi(\cdot \mid O_1, \ldots, O_n)$ specified in (9), and evaluate the corresponding $\Psi(\hat{P}(E_b))$, $b = 1, \ldots, B$, which gives us a random sample from the posterior distribution of the survival function at x_0, given the observed data $O_1, \ldots, O_n$.

Properties of targeted posterior distribution of survival function and comparison with standard Bayesian learning: A standard Bayesian approach would involve specifying a parametric model, specifying a prior distribution on all the parameters of this parametric model, calculating the corresponding posterior distribution involving sampling from a high dimensional multivariate density (since there are many parameters), and model selection (e.g.) based on the posterior density so that these calculations will have to be carried out for lots of candidate parametric models. In spite of the computational challenges and effor of this standard Bayesian approach, the resulting estimator of the survival function will typically be too biased due to model miss-specification. Model selection using a likelihood or Bayesian criteria would generally not reduce the bias at the wished rate of $o(1/\sqrt{n})$, since the selection is in essence based on a bias variance trade off for the purpose of estimating the whole density. As a consequence, the relative efficiency of the simple empirical survival probability and such a standard Bayesian estimator (e.g posterior mean) would converge to infinity in favor of the empirical survival function. The same criticism would apply to a sieve based (frequentist) maximum likelihood estimator using (say) likelihood based cross-validation to select models or other fine tuning parameters. The problem of both the Bayesian and maximum likelihood estimation methodology is that the estimation and model selection are not targeted towards the nice smooth parameter being the survival probability, so that the resulting estimation procedure involves the wrong bias variance trade off.

On the other hand, the targeted empirical posterior Bayesian distribution is centered at the efficient empirical survival probability (recall $\Psi(\hat{P}) = 1/n \sum_i I(O_i > x_0)$), and the spread of the posterior distribution is asymptotically completely driven by the variance of this efficient empirical survival function estimate (and by the prior distribution for small samples). In addition, the calculations for establishing this targeted posterior distribution only involve sampling from a univariate posterior density and is therefore easy and fast from a computational point of view.

24 Generalizations of targeted MLE to general loss functions.

The targeted MLE methodology is a variation of the general method of modifying a current estimate P^0 of the true probability distribution P_0 in a new estimate $P^* = P^*(P^0)$ of the true probability distribution in such a way that it "targets" a particular parameter of the true probability distribution in a particular model, where this "targeting" is formalized by requiring that the updated estimate solves the efficient influence curve based estimating equation, and preferably it also increases the likelihood relative to P^0: i.e., if $D^*(P)$ is the efficient influence curve of the parameter at a P in the model $\mathcal{M}$ for all P, then we require for the update P^* that

$$0 = \sum_{i=1}^{n} D^*(P^*)(O_i).$$

We showed how this can be achieved by either defining an ϵ-extension $P^0(\epsilon)$ and finding an ϵ_n solving $0 = \sum_i D^*(P^0(\epsilon_n))(O_i)$, or, 2) by requiring that the model $\{P^0(\epsilon) : \epsilon\}$ has score at $\epsilon = 0$ contained in the linear span of the efficient influence curve $D^*(P^0)$ and iteratively maximizing the likelihood over ϵ while updating the current estimate accordingly till convergence.

The fact that P^* solves the efficient influence curve equation allows one to establish that the corresponding parameter $\Psi(P^*)$ of P^* is a double robust and locally efficient estimator of $\psi_0 = \Psi(P_0)$ in censored and causal inference models and general data structures, under similar conditions as required for the analysis of solutions of optimal estimating equations (van der Laan, Robins, 2003). A very nice advantage of targeted MLE relative to the estimating equation approach in (van der Laan, Robins, 2003) is that we can still evaluate the performance of the updated $P^*(P^0)$ by its log-likelihood value, so that we can easily handle selection of the initial P^0, selection of the ϵ-fluctuation model among a set of such ϵ-fluctuations, and/or multiple solutions for ϵ_n. In addition, having the log-likelihood criteria also allows us to nicely generalize the targeted maximum likelihood method for estimation of path-wise differentiable parameters to non-pathwise differentiable parameters as in van der Laan, Rubin (2006).

Although we focussed in our presentation of the targeted ML methodology and in the examples on using the log-likelihood as criteria, other criteria than the log-likelihood can be used as well, while still preserving the statistical properties of the resulting substitution estimator $\Psi(P^*)$ for the parameter of interests. In fact, one can use any loss function $L(P)(O)$ for the true proba-

bility distribution P_0 satisfying that $\arg\min_{P\in\mathcal{M}} E_0 L(P)(O)$ (which can have lots of solutions) uniquely identifies $D^*(P_0)$. Thus, if $D^*(P_0)$ only depends on "components" of the probability distribution P_0, then the loss function $L(P)$ is allowed to also only use these components of P. In this case, we should call the methodology *targeted minimum loss learning* instead of targeted maximum likelihood learning. In this case, given an initial P^0, one finds an update P^* with a preferably improved loss (i.e, $\sum_i L(P^*)(O_i) \geq \sum_i L(P^0)(O_i)$) solving $0 = \sum_i D^*(P^*)(O_i)$. Our iterative targeted maximum likelihood can also be generalized by selecting an ϵ-fluctuation, $\{P^k(\epsilon) : \epsilon\}$, for each current fit P^k, satisfying

$$\frac{d}{d\epsilon} \sum_i L(P^k(\epsilon))(O_i)\Big|_{\epsilon=0} = \sum_i D^*(P^k)(O_i).$$

By our proofs it follows that by iteratively carrying out the update $P^k(\epsilon_n^k)$ with

$$\epsilon_n^k = \arg\min_{\epsilon} \sum_i L(P^k(\epsilon))(O_i),$$

it follows that we converge to a solution of the efficient influence curve equation, while also increasing the performance w.r.t. to the empirical loss.

Another generalization is obtained by replacing the efficient influence curve by an alternative inefficient influence curve (i.e., a gradient of the path-wise derivative). In other words, in the above one can replace the efficient influence curve $D^*(P)$ at P by any influence curve $D(P)$ at P. In this case, one finds an update $P^* = P^*(P^0)$ of P^0 so that $0 = \sum_i D(P^*)(O_i)$, and preferably this update improves the empirical loss. The ϵ-fluctuation would now also be based on more ad hoc ϵ-fluctuations, but still allowing for an improved empirical loss. The resulting substitution estimator $\Psi(P^*)$ of $\psi_0 = \Psi(P_0)$ will still have the (e.g., double) robustness and asymptotic linearity properties, but it will not be fully efficient.

25 The inclusion of data adaptive regression methodology into targeted maximum likelihood learning.

In this section we discuss different methods for targeted maximum likelihood learning that include machine learning to fit the nuisance parameters. In order to carefully define different proposals we will consider the semi-parametric regression problem as example. After having presented the various approaches,

we will aim to compare them w.r.t their properties and make a decision for a favorite proposal.

So we observe $O_i = (W_i, A_i, Y_i)$, $i = 1, \ldots, n$, and we assume the model $E_0(Y \mid A = a, W) - E_0(Y \mid A = 0, W) = m(a, W|\beta_0) = A(\beta_0^\top V)$ for a $V \subset W$. Let β_0 be the parameter of interest. Here W will represent a user supplied adjustment set that can thus be a subset of the total set of baseline covariates, but in order to avoid unnecessary notation we will just denote this target adjustment set with W.

The targeted MLE of β_0 is an estimator $P_n \to \hat{\Psi}(P_n)$ that is indexed by the initial estimator of $Q_{\beta_0,\theta_0}(A, W) = E_0(Y \mid A, W) = m(A, W \mid \beta_0) + \theta_0(W)$, where $\theta_0(W) = E_0(Y \mid A = 0, W)$, and $\Pi_0(W) = E_0(A \mid W)$: i.e., the TMLE requires an initial density estimator of the distribution P_0 of $O = (W, A, Y)$. Specifically, let $\hat{\Gamma}_1(P_n)$ be an initial estimator of $Q_0(A, W)$, and let $\hat{\Gamma}_2(P_n)$ be an estimator of $\Pi_0(W)$, so that we can denote these two nuisance parameters and estimator with γ_0 and $\hat{\Gamma}(P_n)$, respectively.

Given an estimator $\hat{\Gamma}(P_n)$, the targeted maximum likelihood estimator is obtained by iterating the following procedure:

Initial estimators: $\hat{Q}^0$, $\hat{\Pi}^0$ Let $\hat{\beta}^0$, and $\hat{\theta}^0(P_n)$ be initial estimators of β_0 and $\theta_0(W)$ based on $\hat{\Gamma}_1(P_n)$. Let $\hat{\Pi}^0(P_n) = \hat{\Gamma}_2(P_n)$ be the initial estimator of $E(A|W)$.

Orthogonalize A: Let $\hat{P}i^1(P_n)$ be an update of $\hat{\Pi}^0(P_n)$ obtained by adding covariate $\epsilon(\hat{\theta}^0(P_n)(W)V_j : j)$ extension, where ϵ is fitted with MLE, so that $P_n(A - \hat{\Pi}^1(W))V_j(\hat{\theta}^0(W)) = 0$, $j = 1, \ldots, J$. Thus, $A - E^1(A \mid W)$ is empirically uncorrelated with $V_j\hat{\theta}^0(W)$ for all V_j in model $A(\beta V)$.

Compute first step TMLE-update: Compute first step targeted MLE update $\hat{Q}^1$, or equivalently, $\hat{\beta}^1$ and $\hat{\theta}^1$, by regressing $m(A, V \mid \hat{\beta}^0) + \hat{\theta}^0(P_n)(W) + \epsilon(A - \hat{\Pi}^1(W))V$.

Iterate till convergence: Go back to the orthogonalization step and TMLE-update step with this new choice $\hat{Q}^1$, $\hat{\Pi}^1$, and iterate carrying out these two steps till convergence of $\hat{Q}^k, \hat{\Pi}^k$.

Therefore, to completely define a targeted MLE estimator $\hat{\Psi}(P_n)$ it remains to define the initial estimator $\hat{Q}^0$ and $\hat{\Pi}^0$. We wish to use data adaptive regression/estimation methodology to estimate these important nuisance parameters, which represents a data adaptive way to deal with confounding by W. Our estimator will be double robust.

25.1 Data adaptive estimation of $\Pi_0(W)$.

Firstly we fit $\Pi^0(W)$ by using the squared error loss function

$$\Pi_0 = \arg\min_{\Pi} E_0(A - \Pi(W))^2.$$

In particular, we can use a super learner that involves proposing a set of candidate learners $\hat{\Pi}_j$ of Π_0 and minimizing

$$\alpha \to \sum_v \sum_{i \in V(v)} (A_i - \sum_j \alpha(j)\hat{\Pi}_{jv}(W_i))^2, \tag{10}$$

where $\hat{\Pi}_{jv}$ is the j-th learner $\hat{\Pi}_j$ applied to the v-th learning sample P^1_{nv}, and $V(v)$ is the v-th validation sample. Let α_n be the minimizer so that $\sum_j \alpha_n(j)\hat{\Pi}_j(P_n)$ is the super learner. We propose various ways of shrinking this super learner to control its potential overfitting.

Select among candidate learners: Firstly, we propose to rank the candidates by cross-validated risk, and run a linear regression of A_i on the top k candidates to determine the super learner for these top k learners:

$$\alpha \to \sum_v \sum_{i \in V(v)} (A_i - \sum_{j=1}^{k} \alpha(j)\hat{\Pi}_{(j)v}(W_i))^2,$$

and select k based on BIC. Thus, the next best learner is only added to the super learner if it improves the cross-validated risk by more than the penalty induced by BIC for adding an extra parameter to a model.

Convex combinations only: We also suggest that it is a good idea to restrict the linear combinations of candidate learners in the super learner to convex combinations. So α_n is defined as a minimizer of (10) over $\{\alpha : 1 \geq \alpha(j) \geq 0, \sum_j \alpha(j) \leq 1\}$ or we could restrict to $\alpha(j)$ with $\sum_j \alpha(j) = 1$.

Double cross-validation to select among super learner and other candidates: We can define the set of candidate learners as $\hat{\Pi}_j(P_n)$, $j = 1, \ldots, J$, augmented with the super learner $\hat{\Pi}(P_n) = \sum_j \alpha_n(j)\hat{\Pi}(P_n)$, and select among these $J+1$ candidates the one that minimizes cross-validated risk. This requires double cross-validation so that it is recommended to select V large, such as 10 or 20. One could also include here the super learner based on the top k candidate learners as additional candidates and thus use honest cross-validation to select among the candidate learners, and the super learner based on the top k candidate learners, $k = 1, \ldots, J$.

25.2 (Targeted) squared error loss function for Q_0, given fit of Π_0.

For fitting $Q_0(A,W) = A(\beta_0 V) + \theta_0(W)$ we use the squared error loss function possibly with weights depending on $\Pi^0(W)$:

$$Q_0 = \arg\min_Q E_0(Y - Q(A,W))^2 w(\Pi^0)(A,W)).$$

In particular, these weights could be inspired directly by the efficient influence curve of β_0:

$$\begin{aligned} IC^*(P_0)(O) &= \left\{c_0^{-1}(A - E_0(A \mid W))V\right\}(Y - Q_{\beta_0,\theta_0}(A,W)) \\ &= (A - E_0(A \mid W))(Y - Q_{\beta_0,\theta_0}(A,W))c_0^{-1}V \end{aligned}$$

Here c_0 is the standardization matrix obtained by differentiating the estimating function w.r.t. β:

$$c_0 = -\frac{d}{d\beta_0} E_0(A - E_0(A \mid W))V(Y - Q_{\beta_0,\theta_0}(A,W)).$$

Thus,

$$IC^*(P_0)IC^{*\top}(P_0) = c_0^{-1}VV^\top c_0^{-1\top}(A - E_0(A \mid W))^2(Y - Q_{\beta_0,\theta_0}(A,W))^2.$$

Since we wish to minimize variance of the efficient influence curve components, this expression suggests the following weighted squared error loss function:

$$L_{w(\Pi)}(O,Q) = (A - \Pi(W))^2(Y - Q(A,W))^2,$$

where $w(\Pi) = (A - \Pi(W))^2$ is a weight implied by a fit of Π_0. Thus, to obtain a weighted targeted squared error loss function we simply substitute our data adaptive fit Π^0 to obtain the weights. Even, if Π is misspecified, the loss function $L_{w(\Pi)}(O,Q)$ remains a valid loss function for Q_0. The advantage of this weighted loss function relative to the regular squared error loss function corresponding with $w = 1$ is that it aims to minimize the variance of the parameter of interest while it still is concerned iwth the overall fit of Q_0 as well.

25.3 Super learning of Q_0 given $\hat{\Pi}^0$:

Given this targeted squared error loss function, we can run any type of machine learning algorithm including our super learning approach based on candidate

fits (which could be the initial estimators or the actual updated targeted MLE corresponding with these initial estimators) of Q_0. For example, given candidate fits $\hat{Q}_j(P_n)$ of Q_0 of the form $A(\hat{\beta}_j V)+\hat{\theta}_j(W)$, $j=1,\ldots,J$, we can define the super learner as $\sum_j \alpha_n(j)\hat{Q}_j(P_n)$ where

$$\alpha_n = \arg\min_{\alpha} \sum_v \sum_{i \in V(v)} w_i(\Pi^0)(Y_i - \sum_j \alpha(j)\hat{Q}_j(P^0_{nv})(A_i,W_i))^2,$$

and P^0_{nv} denotes the empirical distribution of the v-th training sample.

For example, the candidate learners $\hat{Q}_j(P_n)$ could be a DSA algorithm, MARS, or linear main term regression algorithms such as LARS always forcing in the model terms AV_j.

25.4 Candidate fits that are concerned about lack of adjustment due to forced inclusion of A-terms.

The forced inclusion of the $m(A,V \mid \beta)$ component makes it harder for terms that are correlated with A to be selected by the particular regression algorithm concerned with fitting $\theta_0(W)$. For the assessment of the effect of A controlling for W this is a serious concern. In this subsection we consider approaches that aims to select variables taht would also have been selected if the A-terms would not have been present.

Consider the following approach: First apply an arbitrary data adaptive regression of Y on A,W (e.g., MARS, Random Forest, Neural Networks), then evaluate this data adaptive fit at $A=0$ to obtain a fit of $\theta_0(W)$, and finally fit a linear regression in main terms AV_j using as off-set the just obtained fit of $E(Y \mid A=0,W)$. The advantage of such fits, relative to fits that force in the AV_j terms from the start, is that the terms AV_j are added afterwards so that confounders W that are correlated with A (and AV_j) have a fair chance to be included in the model, which is important for assessing the causal effect of A. A disadvantage of such fits is that this approach results in slight overfits of Q_0, due to the use of internal cross-validation in the machine learning algorithms that does not take into account the additional terms AV_j to be added afterwards. In the following paragraph we argue that this type of overfitting might be handled by the super learning that allows shrinkage.

Shrinking slight overfits based on cross-validation: Even though the latter types of fit could result in overfits of Q_0, due to the use of internal cross-validation in the machine learning algorithms that does not take into account the additional terms AV_j to be added, at the super learning step such fits can be shrunk by the selection of α_j, and might thereby be very appropriate

for our goal. For example, consider such a data adaptive regression fit of $A(\beta_0 V) + \theta_0(W)$ and denote it with $\hat{Q}_j(P_n)$. One could now define candidate learners as $\alpha(j)\hat{Q}_j(P_n)$ and select the shrinkage constant $\alpha(j)$ as

$$\alpha_n(j) = \arg\min \sum_v \sum_{i \in V(v)} \left\{ Y_i - \alpha(j)\hat{Q}_j(P_{nv}^0)(W_i, A_i) \right\}^2,$$

whose solution exists in closed form and is given by:

$$\alpha_n(j) = \frac{\sum_v \sum_{i \in V(v)} Y_i \hat{Q}_j(P_{nv}^0)(W_i, A_i)}{\sum_v \sum_{i \in V(v)} \left\{ \hat{Q}_j(P_{nv}^0)(W_i, A_i) \right\}^2}.$$

Note that indeed, if the candidate learner is weakly performing as a predictor due to being an overfit the numerator will be small relative to the denominator resulting in shrinkage of the predictor. In general, by using the super learner based on various candidate learners, this shrinkage will be applied in a multivariate context.

Other approaches for estimation of $E_0(Y \mid A, W)$: First fit $E(Y \mid W)$. In the above algorithm one runs a regression of Y on A, W in which no preference is given to the A term (e.g., it might not be selected), and subsequently, one adds to the resulting $E(Y \mid A = 0, W)$-fit the parametric component $m(A, V \mid \beta)$. There is still a concern that A might be selected early on in the algorithm and thereby obstruct inclusion of important confounders that are strongly correlated with A. So we now wish to consider approaches that completely leave out A at the start by first focussing on fitting $E(Y \mid W)$. Firstly, we note that given a model $E(Y \mid A, W) = m(A, V \mid \beta_0) + \theta_0(W)$ with $m(0, V \mid \beta_0) = 0$ and $\theta_0(W) = E(Y \mid A = 0, W)$, we have

$$E(Y \mid W) = E(m(A, V \mid \beta_0) \mid W) + \theta_0(W)$$

so that

$$\theta_0(W) = E(Y \mid W) - E(m(A, V \mid \beta_0) \mid W).$$

In other words, if m is linear in A, then one can represent

$$Q_0(A, W) = m(A - E_0(A \mid W), V \mid \beta_0) + E_0(Y \mid W).$$

This suggests the following estimator of Q_0: given a fit $\Pi^0(W)$ of $E_0(A \mid W)$, a fit $Q^0(W)$ of $E(Y \mid W)$, fit β in the model $m(A - \Pi^0(W), V \mid \beta) + Q^0(W)$, and let $Q^0(A, W) = m(A - \Pi^0(W), V \mid \beta_n) + Q^0(W)$.

A nice advantage of this approach is that one can use super learning to fit $E(Y \mid W)$ ignoring A, and subsequently carry out targeted further adjustment

through a fit of $E(A \mid W)$ to adjust the fit $E(Y \mid W)$ into a fit of $E(Y \mid A = 0, W)$. The disadvantage is that the consistency of the resulting fit of $E(Y \mid A = 0, W)$ does now rely on the consistency of the fit of $E(A \mid W)$, so that we loose the double robustness property. On the other hand, we can also think of double robustness in terms of the fact that the bias of the estimating function is a product in the bias of the $E(A \mid W)$ fit and the bias of the $E(Y \mid A = 0, W)$ fit. From that point of view, this second order bias is still preserved, but one cannot only bet on a good fit of $E(Y \mid A = 0, W)$ anymore.

Inspection of bias of two TMLE approaches: To investigate the bias terms for the two approaches, let's consider the case of $m(A, V \mid \beta) = \beta A$. The targeted MLE $\beta_n = \beta_n(\theta_n, \Pi_n)$ based on an initial fit $\beta_n A + \theta_n(W)$ solves the equation $P_n(A - \Pi_n(W))(Y - betaA - \theta_n(W)) = 0$ where Π_n is an estimator $\Pi_0(W) = E_0(A \mid W)$ and $\theta_n(W)$ is an estimator of $E(Y \mid A = 0, W)$. The targeted MLE $\beta_n = \beta_n(\tilde{\theta_n}, \Pi_n)$ based on an initial fit $\beta_n(A - \Pi_n(W)) + \tilde{\theta}_n(W)$ with $\tilde{\theta}_n(W)$ an estimator of $\tilde{\theta}_0(W) = E_0(Y \mid W)$ solves the equation $P_n(A - \Pi_n(W))(Y - \beta(A - \Pi_n(W)) - \tilde{\theta}_n(W)) = 0$. Thus the relevant asymptotic bias terms $\beta(\theta, \Pi) - \beta_0$ and $\beta(\tilde{\theta}, \Pi) - \beta_0$ at fixed (θ, Π) and $(\tilde{\theta}, \Pi)$ follow from the equations $P_0(A - \Pi(W))(Y - \beta A - \theta(W)) = 0$ and and $P_0(A - \Pi(W))(Y - \beta(A - \Pi(W)) - \tilde{\theta}(W)) = 0$, respectively. Simple algebra yields

$$\begin{aligned} \beta(\theta, \Pi) - \beta_0 &= \frac{E_0(\Pi_0 - \Pi)(\theta_0 - \theta)(W)}{E_0(A - \Pi(W))A} \\ \beta(\tilde{\theta}, \Pi) - \beta_0 &= \frac{E_0(-\beta(\Pi - \Pi_0)^2 + (\Pi - \Pi_0)(\tilde{\theta}_0 - \tilde{\theta})}{E_0(A - \Pi)(A - \Pi_0)}. \end{aligned}$$

Note that if both approximations $(\Pi - \Pi_0)$ and $\tilde{\theta} - \tilde{\theta}_0$ are equally effective or Π_0 is easier to fit than $\tilde{\theta}_0$, then the targeted MLE based on initial estimator $\beta_n(A - \Pi_n(W)) + \tilde{\theta}_n(W))$ based on fit $\tilde{\theta}_n(W)$ of $E_0(Y \mid W)$ also results in second order bias. If on the other hand, Π_0 is harder to approximate than θ_0, then the latter targeted MLE is not the preferred method.

Given the semiparametric regression model $Y = A(\beta V) + \theta_0(W)$, it follows that $E(Y \mid W) = \Pi(W)(\beta V) + \theta_0(W)$. Therefore, if one first fits $E(Y \mid W)$, then it makes sense to include as candidate covariates in the regression algorithm the term $\Pi(W)$ or terms $\Pi(W)V_j$. These covariates $\Pi(W)V_j$ should not get a special treatment at this stage but selection is driven by obtaining a good overall fit of $E(Y|A, W)$: these special covariates will be used in the targeted MLE step to obtain the wished bias reduction and thus do not need to obtain twice special treatment.

26 General proposal for TMLE of additive model based variable importance, including model selection on effect modification.

Firstly, we consider the semiparametric regression model $Y = A(\beta_0 V) + \theta_0(W)$ and present our proposal of targeted maximum likelihood estimation of β_0 and θ_0. Subsequently, we show how we can also data adaptively select the model for the effect modification component $A(\beta V)$.

Semiparametric regression model: Assume the model $Y = A(\beta_0 V) + \theta_0(W)$ for user supplied $V \subset W$ and unspecified function $\theta_0(W)$.

Predictor of outcome covariate: Consider an estimator $P_n \rightarrow \hat{\theta}(P_n)$ of $\theta_0(W) = E_0(Y \mid A = 0, W)$ or $E_0(Y \mid W)$. This could be an arbitrary data adaptive estimator such as one based on super learning of Y on A, W and evaluation at $A = 0$, or super learning of Y on W.

Define $W_{1nv} = W_{1nv}(W) = \hat{\theta}(P^0_{nv})(W)$ as the univariate summary measure of W obtained by applying this estimator to the empirical distribution P^0_{nv} of the training sample corresponding with the v-th split in a V-fold cross-validation sample splitting scheme, $v = 1, \ldots, V$, and let $W_{1n} = \hat{\theta}(P_n)(W)$ be the summary measure obtained by applying the estimator to the whole sample.

Propensity score covariate: Consider an estimator $P_n \rightarrow \hat{\Pi}(P_n)$ of $\Pi_0(W) = E_0(A \mid W)$. This could be an arbitrary data adaptive estimator such as one based on super learning.

Define $W_{2n} = W_{2n}(W) = \hat{\theta}(P_n)(W)$ as the univariate summary measure of W obtained by applying this estimator to the empirical distribution P_n.

Reduced Data Structure TMLE: Reduce the observed data to $O = (W_{1ni}, W_{2ni}, V_i, A_i, Y_i)$ and we will now compute the targeted MLE based on this reduced data structure. We will need the covariates on training samples as well to control overfitting and honest evaluation of a regression of Y on these covariates.

As initial estimator of $E(Y \mid A, W)$ we fit a parametric regression of $Y = A(\beta V) + \theta(W_1, W_2, V \mid \alpha)$. We suggest as simple model $\theta(W_1, W_1, V \mid \alpha) = \alpha_0 + \alpha_1 W_1 + \alpha_2 W_2$ or one can also add an interaction $\alpha_2 W_1 W_2$.

We estimate $E(A \mid W)$ with the already available $\Pi_n(W)$. The targeted MLE is now computed based on this initial estimator $Q^0(A,W) = A(\beta_n V) + \theta(W_1, W_2 \mid \alpha_n)$ and $\Pi_n^0 = \Pi_n$, where we can apply the joint updating of both fits Q^k and Π^k, iteratively.

We could use cross-validation to select among different models $\theta(W \mid \alpha)$. For example, consider different estimators $\hat{\theta}_j(P_n)$ of $\theta_0(W)$. One can then select

$$j_n = \arg\min_j \sum_v \sum_{i \in V(v)} (Y_i - A_i \beta(P_{nv}^0) V_i - \hat{\theta}_j(P_{nv}^0)(W_{1nv}(W_i), W_{2n}(W_i)))^2.$$

Note that we substitute the covariate values for W_1 based on the training sample specific transformations $\hat{\theta}(P_{nv}^0)$ of W. In this way we can also run a j-specific data adaptive regression algorithm of Y on (W_{1nv}, W_{2n}, V, A) on the training sample v according to the semiparametric regression model $A\beta V + \theta(W_1, W_2, V)$, for each v, and then fine tune the choice j of this algorithm based on the cross-validation. In particular, one can use cross-validation to decide between the models αW_1, $\alpha(W_1, W_2)$ and $\alpha(W_1, W_2, W_1 W_2)$.

Selection of effect modifiers: Suppose now that we wish to select a model for $E(Y \mid A, W) - E(Y \mid A = 0, W)$. As above, we first compute the reduced data structure $(W_{1n}, W_{2n}, V, A, Y)$ as above, based on estimators of $E(Y \mid A = 0, W)$ (now not based on a model for $E(Y \mid A, W) - E(Y \mid A = 0, W)$) or $E(Y \mid W)$, and $E(A \mid W)$. We now determine a data adaptive estimator of the regression of Y on W_1, W_2, V, A where we use cross-validation to select a final model $Y = m(A, V \mid \beta) + \theta(W \mid \alpha)$. For example, one might run the DSA algorithm indexed by the size of the model k and we would use cros-validation to select k. One can use such a data adaptive regression algorithm to select both functional forms $m(A, V \mid \beta)$ and $\theta(W \mid \alpha)$, or we could assume a fixed model $\alpha(W_1, W_2)$ (say) for θ, and we could also assume a linear form $A\beta V$ for m, so that the model is determined by the selected set of effect modifiers. To evaluate the cross-validated risk of any regression fit we need to know the covariate transformation W_1 on the training sample.

Once the model $m(A, V \mid \beta)$ is selected, we run the targeted maximum likelihood estimator of β for that fixed model corresponding with an initial fit $Q^0 = m(A, V \mid \beta_n) + \theta(W \mid \alpha_n)$ and $\Pi_n^0 = \Pi_n$.

We note that the squared error loss function for Q_0 can here be replaced by the targeted loss function that assigns weights $(A - \Pi(W))^2$.

27 Fitting a marginal structural model without inverse weighting

Let $O = (W, A, Y)$ and suppose we wish to estimate $E(Y(a) \mid V) - E(Y(0) \mid V) = m(a, V \mid \psi_0)$ according to parametric model. First we run a data adaptive targeted maximum likelihood estimator for the parameter $E(Y \mid A = a, W) - E(Y \mid A = 0, W)$, involving model selection on this parameter, according to methods discussed above. Consider this data adaptively selected model as given and denote it with $m(A, W \mid \beta_0)$. So now we work in the model $E(Y \mid A = a, W) - E(Y \mid A = 0, W) = m(a, W \mid \beta_0)$ and we treat ψ_0 as a parameter of β_0, P_0. We can work out the influence curve of the estimator obtained by plugging in an efficient estimator for β_0, which is a simple delta method application. We estimate α by simple substitution of the TMLE of β_0 according to this model, and it follows that that is also the actual TMLE of α_0.

If we apply the statistical inference based on the influence curve implied by this model $m(a, W \mid \beta_0)$, then this does not take into account the model selection on m, so that that can be used as optimistic statistical inference.

This TMLE can now also be used to compute a IPCW-reduced data TMLE for $E(Y(\bar{a}) - Y(\bar{0}) \mid V) = m(\bar{a}, V \mid \psi_0)$ for time-dependent treatment based on longitudinal data structure.

28 Causal effect modification in randomized trial.

In a randomized trial in which one observes $O = (W, A, Y)$, a parameter of interest is

$$\psi_{0j}(w) = E_0 E(Y|A = 1, W(j) = w, W(-j)) - E(Y|A = 0, W(j) = w, W(-j)) \\ - E(Y|A = 1, W(j) = 0, W(-j)) + E(Y|A = 0, W(j) = 0, W(-j)),$$

which measures effect modification by covariate $W(j)$ while controlling for other covariates.

Define $m_0(w, W(-j)) = E(Y|A = 1, W(j) = w, W(-j)) - E(Y \mid A = 0, W(j) = w, W(-j))$. Then

$$\begin{aligned} \psi_{0j}(w) &= m(w, W(-j)) - m(0, W(-j)) \\ &= E(Y|A = 1, w_j, W(-j)) - E(Y|A = 0, w_j, W(-j)) \\ &\quad - E(Y|A = 1, 0, W(-j)) + E(Y|A = 0, 0, W(-j)). \end{aligned}$$

If $W(j)$ is binary, then we can compute the nonparametric efficient influence curve of ψ_{0j} since it equals a simple linear combination of the four parameters $(E(Y(1,1), EY(1,0), EY(0,1), EY(0,0))$. In particular, we can apply the TMLE targeting all four parameters, or direclty ψ_{0j} with a single covariate extension. This would involve adding covariate extensions with covariates like $I(A = 1, W_j = 1)/g(1,1|W)$, and we can factorize the joint treatment mechanism, with one factor being known, but the other is not. So this methodology will require inverse probability of treatment weighting.

Let's now consider a model based approach. Firstly, we assume $E(Y|A_1 = a_1, A_2 = a_2, W) - E(Y|A_1 = 0, A_2 = 0) = m(a_1, a_2, V|\beta_0)$, where A_1 denotes the randomized treatment, and A_2 is the effect modifier $W(j)$. Then our estimator of ψ_{0j} is given by $m(a_1, a_2, V \mid \beta_n) - m(a_1, 0, V \mid \beta_n) - m(0, a_2, V \mid \beta_n) + m(0, 0 \mid \beta_n)$, where β_n is the TMLE. The TMLE involves fitting $E(A_1|W)$ and $E(A_2|W)$, where the first one is known in a randomized trial. For example, consider a model $m(a_1, a_2, V \mid \beta_0) = a_1 m_1(V \mid \beta_0) + a_2 m_2(V \mid \beta_0) + a_1 a_2 m_3(V \mid \beta_0)$. Then, $\psi_{0j}(w) = w m_3(V \mid \beta_0)$.