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

Year 2008 *Paper* 234

Estimation Based on Case-Control Designs with Known Incidence Probability

Mark J. van der Laan*

*Division of Biostatistics and Department of Statistics, University of California, Berkeley, laan@berkeley.edu

http://biostats.bepress.com/ucbbiostat/paper234

Estimation Based on Case-Control Designs with Known Incidence Probability

Mark J. van der Laan

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. The typical approach used in practice is to fit (conditional) logistic regression models, ignoring the case-control sampling, in order to estimate the conditional odds ratios of being a case, given baseline covariates and the exposure of interest. Although these methods do not rely on knowing the true incidence probability (i.e, probability of being a case), and provide valid logistic regression model based estimates of the conditional effect of exposure on odds ratio scale, they do not provide an estimate of a marginal causal odds ratio or causal relative risk, which are causal parameters representing the typical parameters of interest in randomized trials comparing different treatment or exposure levels. By the same argument, these methods do not provide measures of marginal variable importance. In this article we focus on methods for causal inference and variable importance analysis for matched and unmatched case-control studies relying on knowing the incidence probability, conditional on the matching variable if matching is used. We start out with presenting, for both case-control designs, a simple intercept adjustment method that deterministically maps a, possibly weighted for matched case-control designs, logistic regression fit into a valid model based fit of the actual conditional probability on being a case, given the covariates. The resulting estimate of the conditional probability of being a case has now the important property that its standard error is proportional to the incidence probability (divided by the square root of the sample size) so that the obtained

precision is good enough for accurately estimating marginal causal relative risks or causal odds-ratios even when the probability of being a case is extremely rare. Subsequently, we present our general proposed methodology, involving a simple weighting scheme of cases and controls, that 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. For regular case-control designs the weighting only relies on knowing the true population proportion of cases or, equivalenty, the true probability of being a case, and for matched case-control sampling it also relies on knowing this proportion of cases within each population strata of the matching variable. 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. 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. We also illustrate such double robust targeted maximum likelihood estimators in marginal structural models and semi-parametric logistic regression models. Finally, we show that case-control studies nested in randomized trials allow estimation, based on inverse probability of treatment weighted (IPTW) estimators of the marginal causal relative risk or odds ratios without the need to know the incidence probability, and we present the simple implications for observational case-control studies in which this incidence probability is not known but known to be close to zero. By comparing these methods with the efficient method for the case that the incidence probability is known, it follows that even in randomized trials the knowledge of the incidence probability allows for significantly more precise estimation of causal parameters.

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. 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), van der Laan (2006)).

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. Our estimators are not sensitive to the particular choice of coupling. 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.

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 typically 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 Section 8 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 General formulation of case-control sampling.

Above we provided the case control sampling framework for the data structure $O^* = (M, W, A, Y) \sim P_0^*$. In general, we have $O^* = (M, Z, Y) \sim P_0^*$, M the matching variable (which can be chosen to be empty for case control design I), the distribution of interest P_0^* is known to be an element of a model $\mathcal{M}^*$, $\psi_0^* = \Psi^*(P_0^*)$ is a particular parameter of this distribution P_0^* of interest, $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ is a euclidean parameter defined on the model $\mathcal{M}^*$, (M_1, Z_1) is a draw from the conditional distribution of (M, Z), given $Y = 1$, (M_1, Z_2^j) is a draw from the conditional distribution of (M, Z), given $Y = 0, M = M_1$ (or just $Y = 0$ in case control design I), $j = 1, \ldots, J$, and the experimental unit observed data structure for the case-control design is defined as $O = ((M_1, Z_1), ((M_1, Z_2^j) : j = 1, \ldots, J)) \sim P_0$.

The model $\mathcal{M}^*$, and possibly knowing q_0 or $\bar{q}_0(M)$, imply now models for the marginal distribution of (M_1, Z_1) and the marginal distributions of (M_1, Z_2^j), $j = 1, \ldots, J$. The model $\mathcal{M}^*$ does not imply much, if anything,

about the dependence structure among (M_1, Z_1), (M_1, Z_2^j), $j = 1, \ldots, J$, beyond the fact that they share a common variable M_1 for m atched case control studies. Let $\mathcal{M}$ be the model for the observed data distribution P_0 compatible with $\mathcal{M}^*$. One possible and probably the most common model $\mathcal{M}$ is obtained by assuming that, given the first draw (M_1, Z_1) from (M, Z), given $Y = 1$, the control observations are all *independent* draws from the specified conditional distributions. Note that in this latter independence model 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 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 $(M_1, Z_1), (M_1, Z_2^j : j = 1, \ldots, J)$ compatible with its marginal distributions implied by P^*.

In this case Z could include general data structures including censoring or missingness: e..g $Z = (W, \Delta, \Delta A)$ for a missingness variable Δ on the exposure or treatment of interest A. In particular, this general formulation includes that $O^* = (M, L(0), A(0), \ldots, L(K), A(K), Y) \sim P_0^*$ is a longitudinal data structure with a time dependent treatment $\bar{A} = (A(0), \ldots, A(K))$, and $\psi_0^* = \beta_0$ is the unknown parameter of a marginal structural model $E_{P_0^*}(Y_{\bar{a}} \mid V) = m(a, V \mid \beta_0)$, where we have to assume the time-ordering assumption, the consistency assumption and the sequential randomization assumption in causal inference (see e.g. van der Laan and Robins (2002) for an overview of causal inference methods).

Since this generalization is completely straightforward, i.e., just replace (W, A) by a general Z, for the sake of presentation, we focus here on the point treatment data structure in which $Z = (W, A)$ is defined as a set of baseline covariates and a point-treatment/exposure variable of interest.

1.3 Overview of article

In Section 2 we start out with presenting for independent case-control sampling a simple method that deterministically maps the commonly employed logistic regression fit that ignores the case-control sampling into a valid model-based fit of the actual conditional probability on being a case, given the covariates. We extend this methodology to matched case-control designs as well in which case the initial logistic regression fit needs to be based on weighted control observations. For both case-control designs this mapping simply adds an intercept $c_0 = \log q_0/1-q_0$ to the standard or control-weighted

logistic regression fit.

The resulting estimate of the conditional probability of being a case has now the important property that its standard error is proportional to the marginal probability of being a case (divided by the square root of the sample size n) so that the obtained precision is good enough for accurately estimating marginal causal relative risks or causal odds-ratios, *even when the probability of being a case is extremely rare*. We present the formal identifiability results and corresponding method for both matched and unmatched case-control studies.

In Section 3 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. 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 3 and Section 4 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 estimat-

ing 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.

In the subsequent three sections we present various applications of case-control weighted targeted maximum likelihood estimators. In Section 5 we show explicitly (i.e., by example) our general result for case-control design I, that, if ψ_0^* is a marginal causal effect (on the additive, multiplicative or odds ratio scale) and $\mathcal{M}^*$ is the nonparametric model for P_0^*, then, for case-control design I, the case-control weighted efficient influence curve for the parameter ψ_0^* and underlying nonparametric model $\mathcal{M}^*$ equals the wished efficient influence curve of ψ_0^* for the observed data model.

As a consequence of this result, we can 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 method presented in Section 2.

In Section 6 we apply our case-control weighted double robust targeted maximum likelihood and case-control weighted double robust estimating function methodology to estimate causal or variable importance parameters based on assuming a semi-parametric logistic regression model, thereby avoiding the need for inverse weighting by a fit of the treatment mechanism g_0^*.

In Section 7 we apply our case-control weighted targeted maximum likelihood estimator to fit a marginal structural model.

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 $g_0^*(a \mid W) = P_0^*(A = a \mid W)$.

For case-control design I, In Section 8 we address inverse-probability of treatment weighed (IPTW) estimation of the causal relative risk and causal odds ratio based on linear and logistic marginal structural models *in the case that* q_0 *is not known*, but the treatment mechanism g_0^* is known, as it would be in a case-control study which is nested within a randomized trial.

In Section 8, we also show that, if g_0^* is unknown but modelled, and $q_0 \approx 0$, one can estimate g_0^* based on the control observations only, which extends the IPTW estimators to estimators which still do not require knowing q_0. However, since not knowing q_0 and not knowing g_0^* makes these causal parameters non identifiable, we are concerned with the statistical properties of these IPTW- estimators and their potential strong sensitivity to model misspecification for g_0^* so that further (practical) study of this estimator will be needed. That is, these IPTW-estimators target a nonparametrically non-identifiable parameter, which suggests strong sensitivity towards model misspecification for the treatment mechanism. Such sensitivity seems to have been observed in the simulation studies of Mansson et al. (2007) investigating various methods based on the propensity score including the IPTW-estimator for a logistic marginal structural model.

In Section 9, we end this article with a discussion. Various technical proofs are deferred to the Appendix.

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 the internet resulted in a list of 56,000 articles. 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. 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 sam-

pling. 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) and Hosmer and Lemeshow (2000), among others.

The method of adding an intercept to a standard logistic regression fit based on case-control design I, and, in that manner, estimating effects different from the odds-ratio has been presented in the literature (see e.g., Greenland (2004), A.P. Morise (1996)), Wachholder (1996)). The corresponding method presented in this article for matched case-control studies has not been addressed in the literature, as far as we know.

Matched case-control studies can be 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 disease (see, Jewell (2004), Schlesselman (1982)): 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, 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.

2 Identifiability and estimation of causal effects based on logistic regressions.

We present identifiability results of the conditional probability $P_0^*(Y = 1 \mid A, W)$ based on case-control data and knowing q_0 and, for case-control design II, also knowing $\bar{q}_0(M)$. These identifiability results are based on first identifying the conditional odds-ratio, and subsequently mapping this into the wished conditional probability by using q_0. We present these results first for unmatched case-control designs, and subsequently for matched case-control designs.

2.1 Case-Control Design I.

A typical approach for case control studies concerns the use of logistic regression models for $P(Y = 1 \mid A, W)$:

$$Q_0^*(A, W) = P(Y = 1 \mid A, W) = Q_{\beta_0}^*(A, W) = \frac{1}{1 + \exp(-m_{\beta_0}(A, W))}, \quad (1)$$

for some parametrization of $m_\beta(A, W)$ such as $\beta^\top(1, A, W)$.

Let

$$\mathrm{OR}(Q_0^*(a, w)) \equiv \{Q_0^*(a+1, w)/(1 - Q_0^*(a+1, w)\}/\{Q_0^*(a, w)/(1 - Q_0^*(a, w))\}$$

be the Odds-ratio at (a, w) measuring the effect of an increase in A from a to $a + 1$. Due to the identifiability result

$$\begin{aligned} \mathrm{OR}(Q_0^*(a, w)) &= \frac{P_0^*(A = a+1 \mid Y = 1, W = w) P_0^*(A = a \mid Y = 0, W = w)}{P_0^*(A = a+1 \mid Y = 0, W = w) P_0^*(A = a \mid Y = 1, W = w)} \\ &= \frac{P_0(A_1 = a+1 \mid W_1 = w) P_0(A_2 = a \mid W_2 = w)}{P_0(A_2 = a+1 \mid W_2 = w) P_0(A_1 = a \mid W_1 = w)} \end{aligned}$$

it follows that any conditional odds ratio comparing the odds at $A = a + 1$ with the odds at $A = a$ can be identified from the case-control sampling distribution.

For case control design I these odds ratios can be estimated with standard logistic regression ignoring the case control sampling, as is well known. In the next theorem this identifiability result of the odds ratio is stated in terms of our statistical formulation. In addition, the next theorem states that adding an intercept $\log q_0/(1 - q_0)$ to the logistic regression targeted by this method yields the true logistic regression function $P_0^*(Y = 1 \mid A, W)$.

Theorem 1 *Given arbitrary constants c, d_j, $j = 1 \ldots, J$, define*

$$\tilde{Q}_0^* \equiv \arg\max_{Q^*} E_{P_0} c \log Q^*(W_1, A_1) + \frac{1}{J}\sum_{j=1}^{J} d_j \log(1 - Q^*(W_2^j, A_2^j)),$$

where Q^* ranges over all positive functions of (W, A) mapping into $(0, 1)$. Let $Q_0^*(w, a) = P_0^*(Y = 1 \mid W = w, A = a)$. We have

$$\frac{Q_0^*(w,a)}{1 - Q_0^*(w,a)} = \frac{q_0}{1 - q_0}\frac{\bar{d}}{c}\frac{\tilde{Q}_0^*(w,a)}{1 - \tilde{Q}_0^*(w,a)}, \tag{2}$$

where $\bar{d} = 1/J \sum_j d_j$.

If q_0 is known, the identifiability relation (2) immediately implies a corresponding identifiability relation for Q_0^* itself:

$$\begin{aligned} Q_0^* &= Q_{0,q_0}^* \\ &\equiv \frac{c(q_0)\tilde{Q}_0^*/(1 - \tilde{Q}_0^*)}{1 + c(q_0)\tilde{Q}_0^*/(1 - \tilde{Q}_0^*)}, \end{aligned}$$

where $c(q_0) \equiv q_0/(1-q_0)$. Equivalently, if we represent $\tilde{Q}_0^* = 1/(1+\exp(\tilde{h}_0))$, $Q_0^* = 1/(1+\exp(h_0))$ for functions $\tilde{h}_0 = \log \tilde{Q}_0^*/(1-\tilde{Q}_0^*)$ and $h_0 = \log Q_0^*/(1-Q_0^*)$, respectively, then

$$h_0 = \log c_0 + \tilde{h}_0.$$

The resulting identifiability result for (e.g.) $\psi_0(a) \equiv EY_a = E_{P_0^*} Q_0^*(W, a)$ is obtained by averaging Q_{0,q_0}^* over the case control weighted distribution of W, $Q_W^* = q_0 Q_1 + (1 - q_0)Q_0$, where Q_1 is the marginal distribution of W_1 and Q_0 is the marginal distribution of W_2. Thus,

$$\begin{aligned} \psi_0(a) &= \psi_{0,q_0}(a) \\ &\equiv E_0\left\{ q_0 Q_{0,q_0}^*(W_1, a) + \frac{1 - q_0}{J}\sum_j Q_{0,q_0}^*(W_2^j, a) \right\}, \end{aligned}$$

which on its turn maps into an identifiability result for the causal relative risk,

$$\begin{aligned} \psi_{0,RR} &\equiv \frac{\psi_0(1)}{\psi_0(0)} = \frac{\psi_{0,q_0}(1)}{\psi_{0,q_0}(0)} \\ &= \frac{E_0\left\{q_0 Q_{0,q_0}^*(W_1, 1) + (1 - q_0)\frac{1}{J}\sum_j Q_{0,q_0}^*(W_2^j, 1)\right\}}{E_0\left\{q_0 Q_{0,q_0}^*(W_1, 0) + (1 - q_0)\frac{1}{J}\sum_j Q_{0,q_0}^*(W_2^j, 0)\right\}}. \end{aligned}$$

For $q_0 \approx 0$, the case-control weighted distribution of W is well approximated by the distribution of the covariate W for controls, so that this causal relative risk relation is well approximated by

$$\psi_{0,RR} \approx \frac{E_0 \sum_j Q^*_{0,q_0}(W_2^j, 1)}{E_0 \sum_j Q^*_{0,q_0}(W_2^j, 0)}.$$

For the validity of case-control weighting , we refer to the general method of case-control weighting as presented in the next section. **Proof of Theorem 1.** Consider fluctuations $\tilde{Q}^*_0(\epsilon)(Y \mid A, W)$ of $\tilde{Q}^*_0(Y \mid A, W)$ with parameter ϵ and score at $\epsilon = 0$ equal to $h(Y \mid A, W)$ for arbitrary functions $h(Y \mid A, W)$ with mean zero w.r.t. $\tilde{Q}^*_0(Y \mid A, W)$. Since h has mean zero, it follows that $h(0 \mid A, W) = -\frac{\tilde{Q}^*_0}{1-\tilde{Q}^*_0}(A, W)h(1 \mid A, W)$. Substitution of these fluctuations into the log likelihood criterion in Q^* and taking the derivative w.r.t ϵ at $\epsilon = 0$ yields the score functions:

$$0 = E_0 ch(1 \mid W_1, A_1) - \frac{1}{J} \sum_j d_j \frac{\tilde{Q}^*_0}{1 - \tilde{Q}^*_0}(W_2^j, A_2^j) h(1 \mid W_2^j, A_2^j),$$

where $h(1 \mid w, a)$ is now an arbitrary function. The right hand side can be worked out as:

$$0 \;=\; E^*_{A,W} \left\{ ch(1 \mid W, A) \frac{Q^*_0(W, A)}{q_0} - \frac{1}{J} \sum_{j=1}^{J} d_j \frac{\tilde{Q}^*_0}{1 - \tilde{Q}^*_0}(W, A) h(1 \mid W, A) \frac{1 - Q^*_0(W, A)}{1 - q_0} \right\}.$$

Since this equation needs to hold for all functions $h(1 \mid W, A)$ it follows that

$$c \frac{Q^*_0(W, A)}{q_0} - \bar{d} \frac{\tilde{Q}^*_0}{1 - \tilde{Q}^*_0}(W, A) \frac{1 - Q^*_0(W, A)}{1 - q_0} = 0$$

or equivalently,

$$\frac{Q^*_0}{1 - Q^*_0} = \frac{\bar{d}}{c} \frac{q_0}{1 - q_0} \frac{\tilde{Q}^*_0}{1 - \tilde{Q}^*_0}.$$

□

Estimation of causal effects based on logistic regression.

This teaches us that we can define (setting, $c = 1$, $\bar{d} = 1$ in the theorem)

$$\tilde{Q}^*_n \equiv \arg\max_\beta \sum_{i=1}^{n} \log Q^*_\beta(W_{1i}, A_{1i}) + \frac{1}{J} \sum_{j=1}^{J} \log(1 - Q^*_\beta(W_{2i}^j, A_{2i}^j)),$$

which can be computed with standard logistic regression software.

Let $\tilde{h}_n = \log \tilde{Q}_n^*/(1-\tilde{Q}_n^*)$ be the log-odds of $\tilde{Q}_n^*$. In addition, one can use this standard log likelihood loss function to carry out model selection based on cross-validation and one can apply data adaptive logistic regression algorithms.

Clearly, the variance of the resulting estimator of the odds ratio $OR(Q_0^*(w,a))$ at a particular w, a does not suffer from the singularity $q_0 \approx 0$. In addition, if q_0 is known, the identifiability relation (2) immediately implies a corresponding estimator of Q_0^* given by

$$Q_{n,q_0}^* \equiv c(q_0)\frac{\tilde{Q}_n^*/(1-\tilde{Q}_n^*)}{1+c(q_0)\tilde{Q}_n^*/(1-\tilde{Q}_n^*)},$$

which is equivalent with adding an intercept $\log c_0$ to the log odds fit $\tilde{h}_n = \log \tilde{Q}_n^*/(1-\tilde{Q}_n^*)$:

$$h_n \equiv \log Q_{n,q_0}^*/(1-Q_{n,q_0}^*) = \log c_0 + \tilde{h}_n.$$

Since the standard error of this estimator Q_{n,q_0}^* is proportional to c_0 and thus q_0, this estimator will result in stable estimators of the causal relative risk or causal odds ratio not suffering from the typical singularity $q_0 \approx 0$.

The resulting estimator of EY_a is obtained by averaging Q_{n,q_0}^* over the case-control weighted empirical distribution of W, and is thus given by

$$\psi_{n,q_0}(a) = \frac{1}{n}\sum_{i=1}^n q_0 Q_{n,q_0}^*(W_{1i},a) + (1-q_0)\frac{1}{J}\sum_j Q_{n,q_0}^*(W_{2i}^j,a),$$

which now maps into an estimator of the causal relative risk,

$$\begin{aligned}\psi_{n,RR} &= \frac{\psi_{n,q_0}(1)}{\psi_{n,q_0}(0)} \\ &= \frac{\frac{1}{n}\sum_{i=1}^n q_0 Q_{n,q_0}^*(W_{1i},1) + (1-q_0)\frac{1}{J}\sum_j Q_{n,q_0}^*(W_{2i}^j,1)}{\frac{1}{n}\sum_{i=1}^n q_0 Q_{n,q_0}^*(W_{1i},0) + (1-q_0)\frac{1}{J}\sum_j Q_{n,q_0}^*(W_{2i}^j,0)}\end{aligned}$$

For $q_0 \approx 0$, the case control weighted empirical distribution of W is well approximated by the pooled empirical distribution of the controls, so that this causal relative risk estimator is well approximated by (for variable J one replaces J by J_i)

$$\psi_{n,RR} \approx \frac{\sum_{i=1}^n \frac{1}{J}\sum_j Q_{n,q_0}^*(W_{2i}^j,1)}{\sum_{i=1}^n \frac{1}{J}\sum_j Q_{n,q_0}^*(W_{2i}^j,0)}$$

It is important to note that without knowing q_0 it would not have been possible to map the standard logistic regression fit $\tilde{Q}_n^*$, that ignores the case-control sampling and yields a robust (against $q_0 \approx 0$) and consistent estimator of the odds ratio of Q_0^*, into an estimator of EY_a which has a standard error proportional to q_0, and thereby in a robust (against $q_0 \approx 0$) estimator of the causal relative risk or causal odds ratio.

Notation: We introduce now some useful notation. Given a function $D^*(O^*)$, we define $P_{0,q_0}D^* = P_0 D_{q_0}$, where $D_{q_0}(O) \equiv q_0 D^*(W_1, A_1, 1) + \frac{1}{J}\sum_{j=1}^J \bar{q}_0(M_1)D^*(W_2^j, A_2^j, 0)$. Similarly, we define $P_{n,q_0}D^* = P_n D_{q_0}$, 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 $1 - q_0$. We refer to D_{q_0} as the case-control weighted version of D^*.

2.2 Matched case-control design II.

Consider a logistic regression model $m_\beta(A, W)$ (1) for $P_0^*(Y = 1 \mid A, W)$. For the matched case control design II the odds ratios can be estimated with standard logistic regression assigning the weights 1 to the cases and weighting the controls by $\bar{q}_0(M_1)$, as we show in the next theorem. In the next theorem we also show how this odds ratio can be mapped into $P_0^*(Y = 1 \mid A, W)$ by adding the intercept $\log c_0$, as in the previous Theorem 1.

Theorem 2 *Given arbitrary constants c, d_j, $j = 1 \ldots, J$, define*

$$\tilde{Q}_0^* \equiv \arg\max_{Q^*} E_0 c \log Q^*(M_1, W_1, A_1) + \bar{q}_0(M_1)\frac{1}{J}\sum_{j=1}^J d_j \log(1 - Q^*(M_1, W_2^j, A_2^j)),$$

where Q^* varies over positive valued functions mapping into $(0, 1)$.

Let $Q_0^*(a, w) = P_0^*(Y = 1 \mid A = a, W = w)$. We have

$$\frac{Q_0^*(m, w, a)}{1 - Q_0^*(m, w, a)} = q_0 \frac{\bar{d}}{c} \frac{\tilde{Q}_0^*(m, w, a)}{1 - \tilde{Q}_0^*(m, w, a)}, \tag{3}$$

where $\bar{d} = 1/J \sum_j d_j$.

For $c = q_0$, this implies $Q_0^* = \tilde{Q}_0^*$.

Equivalently, one adds an intercept $\log q_0 \bar{d}/c$ to the log odds $\tilde{h}_0 = \log \tilde{Q}_0^*/(1 - \tilde{Q}_0^*)$ of $\tilde{Q}_0^*$:

$$h_0 \equiv \log Q_{0,q_0}^*/(1 - Q_{0,q_0}^*) = \log q_0 \bar{d}/c + \tilde{h}_0.$$

The resulting identifiability result for $\psi_0(a) = EY_a = E_{P_0^*}Q_0^*(M,W,a)$ *is obtained by averaging* Q^*_{0,q_0} *over the case control weighted distribution of* M,W*,* $Q^*_{M,W} = q_0Q_1 + \bar{q}_0Q_0$*, where* Q_1 *is the marginal distribution of* (M_1,W_1) *and* Q_0 *is the marginal distribution of* (M_1,W_2)*:*

$$\begin{aligned}\psi_0(a) &= \psi_{0,q_0}(a) \\ &= E_0\left\{q_0Q^*_{0,q_0}(M_1,W_1,a) + \bar{q}_0(M_1)\frac{1}{J}\sum_j Q^*_{0,q_0}(M_1,W_2^j,a)\right\}. \quad (4)\end{aligned}$$

This implies an identifiability result for the causal relative risk,

$$\psi_{0,RR} = \frac{\psi_{0,q_0}(1)}{\psi_{0,q_0}(0)} = \frac{E_0\left\{q_0Q^*_{0,q_0}(M_1,W_1,1) + \bar{q}_0(M_1)\frac{1}{J}\sum_j Q^*_{0,q_0}(M_1,W_2^j,1)\right\}}{E_0\left\{q_0Q^*_{0,q_0}(M_1,W_1,0) + \bar{q}_0(M_1)\frac{1}{J}\sum_j Q^*_{0,q_0}(M_1,W_2^j,0)\right\}}.$$

For $q_0 \approx 0$, the case control weighted distribution of M,W is well approximated by Q_0, i.e. the distribution of the covariate W for controls, so that this causal relative risk relation is well approximated by

$$\psi_{0,RR} \approx \frac{E_0\frac{1}{J}\sum_j Q^*_{0,q_0}(M_1,W_2^j,1)}{E_0\frac{1}{J}\sum_j Q^*_{0,q_0}(M_1,W_2^j,0)}$$

Proof of Theorem 2. Consider fluctuations $\tilde{Q}_0^*(\epsilon)(Y \mid A,M,W)$ of $\tilde{Q}_0^*(Y \mid A,M,W)$ with parameter ϵ with score at $\epsilon = 0$ equal to $h(Y \mid A,M,W)$ for arbitrary functions $h(Y \mid A,M,W)$ with mean zero w.r.t. $\tilde{Q}_0^*(Y \mid M,A,W)$. Note that $h(0 \mid A,M,W) = -\frac{\tilde{Q}_0^*}{1-\tilde{Q}_0^*}(A,M,W)h(1 \mid A,W)$. Substitution of these fluctuations into the log likelihood criterion in Q^* and taking the derivative w.r.t ϵ at $\epsilon = 0$ yields the score functions:

$$0 = E_0 ch(1 \mid M_1,W_1,A_1) - \frac{1}{J}\sum_j d_j\bar{q}_0(M_1)\frac{\tilde{Q}_0^*}{1-\tilde{Q}_0^*}(M_1,W_2^j,A_2^j)h(1 \mid W_2^j,A_2^j),$$

where $h(1 \mid w,a)$ is now an arbitrary function. The right hand side can be worked out as:

$$\begin{aligned}&0 = E^*_{A,M,W}\left\{ch(1 \mid M,W,A)\tfrac{Q_0^*(M,W,A)}{q_0}\right\} \\ &- \textstyle\int_{a,m,w}\frac{1}{J}\sum_{j=1}^J d_j\frac{\tilde{Q}_0^*}{1-\tilde{Q}_0^*}(m,w,a)h(1 \mid m,w,a)\bar{q}_0(m)P_{M_1}(m)P_0^*(w,a \mid Y=0,m) \\ &= E^*_{A,M,W}\left\{ch(1 \mid M,W,A)\tfrac{Q_0^*(M,W,A)}{q_0}\right\} \\ &- \textstyle\int_{a,m,w}\frac{1}{J}\sum_{j=1}^J d_j\frac{\tilde{Q}_0^*}{1-\tilde{Q}_0^*}(m,w,a)h(1 \mid m,w,a)P(M=m,W=w,A=a,Y=0) \\ &= E^*_{A,M,W}\left\{h(1 \mid M,W,A)\left\{c\tfrac{Q_0^*(M,W,A)}{q_0} - \bar{d}\tfrac{\tilde{Q}_0^*}{1-\tilde{Q}_0^*}(M,W,A)(1-Q_0^*(M,W,A))\right\}\right\}.\end{aligned}$$

Since this equation needs to hold for all functions $h(1 \mid W, A)$ it follows that

$$c\frac{Q_0^*(M,W,A)}{q_0} - \bar{d}\frac{\tilde{Q}_0^*}{1-\tilde{Q}_0^*}(M,W,A)(1-Q_0^*(M,W,A)) = 0,$$

or equivalently,

$$\frac{Q_0^*}{1-Q_0^*} = \frac{\bar{d}}{c}q_0\frac{\tilde{Q}_0^*}{1-\tilde{Q}_0^*}.$$

□

Estimation of marginal causal effects based on a logistic regression fit.

This teaches us that we can define (e.g., $c = 1$, $\bar{d} = 1$)

$$\tilde{Q}_n^* \equiv \arg\max_\beta \sum_{i=1}^n \log Q_\beta^*(M_{1i}, W_{1i}, A_{1i}) + \bar{q}_0(M_{1i})\frac{1}{J}\sum_{j=1}^J \log(1-Q_\beta^*(M_{1i}, W_{2i}^j, A_{2i}^j)),$$

which can be computed with standard logistic regression software using weights for the control observations. In addition, one can use this log likelihood loss function to carry out model selection based on cross-validation and one can apply data adaptive logistic regression algorithms. Clearly, the variance of the resulting estimator of the odds ratio $OR(Q_0^*(m,w,a)) = ODDS(Q_0^*(m,w,a+1))/ODDS(Q_0^*(m,w,a))$ at a particular m, w, a does not suffer from the singularity $q_0 \approx 0$.

In addition, the identifiability relation (3) immediately implies a corresponding estimator of Q_0^* itself given by

$$Q_{n,q_0}^* \equiv c(q_0)\frac{\tilde{Q}_n^*/(1-\tilde{Q}_n^*)}{1+c(q_0)\tilde{Q}_n^*/(1-\tilde{Q}_n^*)},$$

or, equivalently, one adds an intercept $\log c_0$ to the log odds fit $\tilde{h}_n = \log \tilde{Q}_n^*/(1-\tilde{Q}_n^*)$:

$$h_n \equiv \log Q_{n,q_0}^*/(1-Q_{n,q_0}^*) = \log c_0 + \tilde{h}_n.$$

Since the standard error of this estimator Q_{n,q_0}^* is proportional to q_0 divided by the square root of the sample size, this estimator will result in stable

estimators of the causal relative risk or causal odds ratio not suffering from the singularity $q_0 \approx 0$.

The resulting estimator of EY_a is obtained by averaging Q^*_{n,q_0} over the case-control weighted empirical distribution of W, and is thus given by

$$\psi_{n,q_0}(a) = \frac{1}{n}\sum_{i=1}^{n} q_0 Q^*_{n,q_0}(M_{1i}, W_{1i}, a) + \bar{q}_0(M_{1i})\frac{1}{J}\sum_j Q^*_{n,q_0}(M_{1i}, W^j_{2i}, a),$$

which now maps into an estimator of the causal relative risk,

$$\begin{aligned}\psi_{n,RR} &= \frac{\psi_{n,q_0}(1)}{\psi_{n,q_0}(0)} \\ &= \frac{\frac{1}{n}\sum_{i=1}^n q_0 Q^*_{n,q_0}(M_{1i}, W_{1i}, 1) + (1-q_0)\frac{1}{J}\sum_j Q^*_{n,q_0}(M_{1i}, W^j_{2i}, 1)}{\frac{1}{n}\sum_{i=1}^n q_0 Q^*_{n,q_0}(M_{1i}, W_{1i}, 0) + (1-q_0)\frac{1}{J}\sum_j Q^*_{n,q_0}(M_{1i}, W^j_{2i}, 0)}\end{aligned}$$

For $q_0 \approx 0$, the case-control weighted empirical distribution of W is well approximated by the empirical distribution Q_{0n} of the controls, so that this causal relative risk estimator is well approximated by

$$\psi_{n,RR} \approx \frac{\frac{1}{n}\sum_{i=1}^n \frac{1}{J}\sum_j Q^*_{n,q_0}(M_{1i}, W^j_{2i}, 1)}{\frac{1}{n}\sum_{i=1}^n \frac{1}{J} Q^*_{n,q_0}(M_{1i}, W^j_{2i}, 0)}$$

3 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*

$$D_{q_0}(O) \equiv q_0 D^*(M_1, W_1, A_1, 1) + \frac{1}{J}\sum_{j=1}^{J} \bar{q}_0(M_1) D^*(M_1, W^j_2, A^j_2, 0),$$

where in the special case of Case Control Design I, we have $\bar{q}_0(M) = 1 - q_0$.

Theorem 3 *(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

$$\bar{q}_0(m) P_0(M_1 = m) P_0^*(W = w, A = a \mid M = m, Y = 0)$$
$$= \bar{q}_0(m) P_0^*(M = m \mid Y = 1) P_0^*(W = w, A = a, M = m, Y = 0) / P_0^*(Y = 0, M = m)$$
$$= P_0^*(M = m, W = w, A = a, Y = 0).$$

This proves that

$$P_0 D = \int_{M_1, W_1, A_1} D(M_1, W_1, A_1, 1) P_0^*(M_1, W_1, A_1, Y = 1)$$
$$+ \frac{1}{J}\sum_{j=1}^{J} \int_{M_1, W_2, A_2} D(M_1, W_2, A_2, 0) P_0^*(M_1, W_2, A_2, Y = 0)$$
$$= P_0^* D = 0.$$

This completes the proof. □

Before we proceed with presenting the statistical implications of this mapping for the analysis of case-control data, we first establish some 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.

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 section 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 4 *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*

$$\frac{d}{d\epsilon}PD_{q_0}(P^*_{S^*}(\epsilon))\Big|_{\epsilon=0} = \frac{d}{d\epsilon}P^*D^*(P^*_{S^*}(\epsilon))\Big|_{\epsilon=0}, \tag{5}$$

where it is assumed that the left and right derivative exist.

By (5) 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 (5) 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 5 *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

$$-\frac{d}{d\epsilon} P D_{q_0}(P^*(\epsilon)) \Big|_{\epsilon=0} = P D_{q_0}(P^*) S(P).$$

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 Preservation of robustness of the 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.

3.3 Case-control weighted estimating functions and locally efficient estimation.

Estimating function methodology developed for prospective sampling immediately implies now, through the case-control weighted mapping, estimating function methodology for case-control sampling.

For sake of presentation, let's start with estimating functions without nuisance parameters. That is, let $\{D_h^*(\psi) : h\}$ denote a class of estimating functions of parameter $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ in model $\mathcal{M}^*$ indexed by functions h ranging over a particular index set. This class maps into a class of case-control weighted estimating functions

$$\left\{ D_{h,q_0}(\psi)(O) = q_0 D_h^*(\psi)(M_1, Z_1, 1) + \bar{q}_0(M_1)\frac{1}{J}\sum_{j=1}^{J} D_h^*(\psi)(M_1, Z_2^j, 0) : h \right\}.$$

Let $\{c_h D_{h,q_0}(\psi_)) : h\}$ be the corresponding set of gradients/influence curves (i.e., the influence curve of the estimator defined as solution of estimating equation implied by D_{h,q_0}), where $c_h = -\frac{d}{d\psi_0} E_0 D_{h,q_0}(\psi_0)^{-1}$. We can now apply (for example) Theorem 2.9 in van der Laan and Robins (2002) to determine the optimal choice h_{opt} of estimating function minimizing $a^\top \Sigma_0(h) a$ for all a, where $\Sigma_0(h)$ denotes the covariance of the influence curve $c_h D_{h.q_0}(\psi_0)$.

This optimal estimating function can now be used to construct locally efficient estimators by defining ψ_n as a solution of $0 = \sum_i D_{h_n,q_0}(\psi)(O_i)$ for an estimator h_n of h_{opt} (or by using corresponding one-step Newton-Raphson estimators), or by constructing a targeted maximum likelihood type estimator P_n^* (van der Laan and Rubin (2006), and see later subsection), and corresponding targeted maximum likelihood estimator $\Psi^*(P_n^*)$ of ψ_0^* solving this equation $0 = \sum_i D_{h(P^*),q_0}(P^*)(O_i)$ viewed as a function in $P^* \in \mathcal{M}^*$, where $h(P^*)$ is a representation of h_{opt} as a function of P^*.

By our Theorems 7 and 8 in the next section it follows that selecting h_{opt} so that $D_{h_{opt}}^*(P^*)$ is the efficient influence curve in model $\mathcal{M}^*$ actually results in the wished optimal choice, and corresponding locally efficient estimating function procedure.

This template for construction of locally efficient estimators can be generalized to estimating functions which also depend on nuisance parameters,

as follows. Let $\{D_h^*(P^*) : h\}$ denote a class of estimating functions of parameter $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$ in model $\mathcal{M}^*$, and let $D_h^*(P^*) = D_h^*(\psi^*, \eta^*)$ for variation independent parameters ψ^* and η^* of P^*. Suppose that these estimating functions are orthogonalized to nuisance parameters in the sense that they satisfy $\frac{d}{d\epsilon} E_0^* D_h^*(P_0^*(\epsilon))\Big|_{\epsilon=0} = 0$ for nuisance fluctuations $P_0^*(\epsilon)$ through P_0^* at $\epsilon = 0$ (i.e. $\Psi^*(P_0^*(\epsilon))$ has derivative zero w.r.t. ϵ at $\epsilon = 0$). In addition, assume that the estimating functions are standardized to have derivative w.r.t. ψ minus the identify matrix:

$$\frac{d}{d\epsilon} E_0^* D_h^*(P_0^*(\epsilon))\Big|_{\epsilon=0} = -\frac{d}{d\epsilon} \Psi^*(P_0^*(\epsilon))\Big|_{\epsilon=0}$$

for fluctuations $P_0^*(\epsilon)$ changing ψ_0^*.

This defines $D_h^*(P^*)$ as a gradient/influence curve of Ψ^* in model $\mathcal{M}^*$ at P^*: see Chapter 1 van der Laan and Robins (2002). This class of gradients/influence curves for model $\mathcal{M}^*$ now maps into the class of case-control weighted functions

$$\left\{ D_{h,q_0}(P^*)(O) = q_0 D_h^*(P^*)(M_1, Z_1, 1) + \bar{q}_0(M_1) \frac{1}{J} \sum_{j=1}^{J} D_h^*(P^*)(M_1, Z_2^j, 0) : h \right\}.$$

Application of Theorem 3 yields that

$$P_0 D_{h,q_0}(P_0^*(\epsilon)) = P_0^* D_h^*(P^*(\epsilon)),$$

so that it follows that also the case-control weighted D_{h,q_0} is an influence curve. Thus this shows that our mapping indeed maps the gradients for parameter Ψ^* for model $\mathcal{M}^*$ into gradients of Ψ^* for model $\mathcal{M}$.

We can now apply Theorem 2.9 in van der Laan and Robins (2002) to determine the optimal choice $h_{opt} = h(P_0^*)$ minimizing $a^\top \Sigma_0(h) a$ for all vectors a, where $\Sigma_0(h)$ denotes the covariance of the influence curve $D_h(P_0^*)$. By our Theorems 7 and 8 it follows that selecting h_{opt} so that $D_{h_{opt}}^*(P^*)$ is the efficient influence curve in model $\mathcal{M}^*$ results in the wished optimal choice. This optimal estimating function $D_{h(P_0)}(\psi_0, \eta_0^*)$ can now be used to construct locally efficient estimators by, given an estimator η_n^*, defining ψ_n as a solution of $0 = \sum_i D_{h_n,q_0}(\psi_n, \eta_n)(O_i)$ or by constructing a targeted maximum likelihood type estimator P_n^* and corresponding substitution estimator $\Psi^*(P_n^*)$ of ψ_0^* solving this equation $0 = \sum_i D_{h(P^*),q_0}(P^*)(O_i)$ viewed as a function in $P^* \in \mathcal{M}^*$.

We note that this approach can be further generalized to estimating functions D_h^* with non-variation independent nuisance parameters.

3.4 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).

The double robust efficient estimating function for sampling from P_0^* is given by

$$\begin{aligned} D^*(\psi^*, g^*, Q^*)(O^*) &= \left\{ \frac{I(A=1)}{g^*(1 \mid M, W)} - \frac{I(A=0)}{g^*(0 \mid M, W)} \right\} (Y - Q^*(M, W, A)) \\ &\quad + Q^*(M, W, 1) - Q^*(M, W, 0) - \psi^*. \end{aligned} \tag{6}$$

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.

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 logistic regression fits as presented in Section 2.

3.5 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 6 *(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 3 and is therefore omitted.

3.6 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.

3.7 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 $D_{q_0}(\psi_0, \eta_0^*)$.

3.8 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

$$\begin{aligned} D^*(g^*, Q^*)(M, W, A, Y) &= \left\{ \frac{I(A=1)}{g^*(1 \mid M, W)} - \frac{I(A=0)}{g^*(0 \mid M, W)} \right\} (Y - Q_2^*(M, W, A)) \\ &\quad + Q_2^*(M, W, 1) - Q_2^*(M, W, 0) - \Psi(Q^*) \\ &\equiv D_1^*(g^*, Q^*)(M, W, A, Y) + D_2^*(Q^*)(M, W), \end{aligned}$$

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}^* presented in Section 2.

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 $\psi_n = \Psi(Q_n^*)$.

4 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.

4.1 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

$$dP(P^*)(W_1, A_1, (W_2^j, A_2^j : j)) = dP^*(W_1, A_1 \mid Y = 1) \prod_{j=1}^{J} dP^*(W_2^j, A_2^j \mid Y = 0), \tag{7}$$

and, for case-control design II, we have

$$\begin{aligned} dP(P^*)(M_1, W_1, A_1, (M_1, W_2^j, A_2^j : j)) &= dP^*(M_1, W_1, A_1 \mid Y = 1) \\ &\quad \prod_{j=1}^{J} dP^*(W_2^j, A_2^j \mid M = M_1, Y = 0). \\ &= dP^*_M(M_1) dP^*(W_1, A_1 \mid M = M_1, Y = 1) \\ &\quad \prod_{j=1}^{J} dP^*(W_2^j, A_2^j \mid M = M_1, Y = 0). \end{aligned} \tag{8}$$

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 (7) or (8), 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 $P(W, A \mid Y = \delta)$.

4.2 Case-control weighting of canonical gradient yields canonical gradient: Case Control Design I.

Firstly, we present the theorem for case-control design I.

Theorem 7 *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 (7), 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 5 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.

4.3 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

$$\psi_n(m) = \frac{\sum_i I(M_{1i} = m) D_{m,q_0}(Q_n^*, g_n^*)(O_i)}{\sum_i I(M_{1i} = m)}, \tag{9}$$

  where

$$D_{m,q_0}(Q^*, g^*)(O) = q_0(1 \mid m) \left\{ \frac{I(A_1 = 1)}{g_0^*(1 \mid m, W_1)} (1 - Q^*(m, W_1, 1)) + Q^*(m, W_1, 1) \right\}$$
$$+ \frac{q_0(0 \mid m)}{J} \left\{ \frac{I(A_2^j = 1)}{g^*(1 \mid m, W_2^j)} (0 - Q^*(m, W_2^j, A_2^j, 1)) + Q^*(m, W_2^j, A_2^j, 1) \right\}.$$

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 (9).

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.

4.4 Case-control weighting of canonical gradient yields canonical gradient: Matched Case Control Design.

For case-control design II, we establish the same result.

Theorem 8 *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 (8), 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 5 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 $\Psi : \mathcal{M} \to \mathbb{R}^d$.

4.5 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 9 *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 $P^*(\epsilon)$ 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 7 or Theorem 8 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.

Proof of Theorems 7 and 8.

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.

5 Double robust locally efficient estimation of marginal causal effects for case-control design I.

5.1 Efficient influence curve of marginal causal effects for case-control design I.

In this subsection we establish that the efficient influence curve of the marginal causal effects defined on a nonparametric model $\mathcal{M}^*$ and indexed by a fixed known q_0 for case-control design I can be represented as a case-control weighted $D_{q_0} = q_0 D^*(\cdot, 1) + (1-q_0) D^*(\cdot, 0)$, with D^* being the efficient influence curve of the marginal causal effect defined on the nonparametric model $\mathcal{M}^*$ (and not indexed by q_0).

Our general theorem 4 and Theorem 7 teach us that this should indeed be true but with D^* being the efficient influence curve of the marginal causal effect defined as $\Psi_{q_0}^*$ indexed by a fixed q_0. Theorem 9 proves that it should also hold for the choice D^* being the efficient influence curve of the marginal causal effect defined as the unrestricted parameter $\Psi^*(P^*) = \Psi_{q(P^*)}^*$. Thus, the result in this subsection is completely predicted by our theorems presented in previous sections.

Theorem 10 *(**Efficient influence curve for case control design I**)*

Consider Case Control Design I with data structure $O = ((W_1, A_1), ((W_2^j, A_2^j) : j))$, *where* (W_1, A_1) *has distribution* $Q_1 \sim (W, A) \mid Y = 1$ *and* (W_2^j, A_2^j) *has*

distribution $Q_0 \sim (W, A) \mid Y = 0$. *Let* $Q_1(w, a) = P(W_1 = w, A_1 = a)$ *and let* $Q_0(w, a) = P(W_2 = w, A_2 = a)$. *Similarly, we define* $Q_1(w) = P(W_1 = w)$ *and* $Q_0(w) = P(W_2 = w)$. *We also define* $Q^*(a, W) = P^*(Y = 1 \mid A = a, W)$, $Q^*_W(w) = P^*(W = w)$. *Define*

$$\begin{aligned}\psi_0 &= f(Q_0^1, Q_1^1, Q_1, Q_0) \\ &\equiv \sum_w \{q_0 Q_1(w) + (1 - q_0)Q_0(w)\} \frac{Q_1(w, 1)q_0}{Q_1(w, 1)q_0 + Q_0(w, 1)(1 - q_0)}.\end{aligned}$$

The efficient influence curve of ψ_0 *is given by*

$$D_{q_0}(P_0)(O) = q_0 D^*(P_0^*)(W_1, A_1, 1) + (1 - q_0)\frac{1}{J}\sum_{j=1}^{J} D^*(P_0^*)(W_2^j, A_2^j, 0)$$

This result teaches us that the variance under P_0 of the weighted double robust estimating function $D_{q_0}(O) = q_0 D^*(W_1, A_1, 1) + \frac{1}{J}\sum_{j=1}^{J}(1-q_0)D^*(W_2^j, A_2^j, 0)$ is the Cramer Rao lower bound for regular asymptotically linear estimators of $\Psi_{q_0}(P_0)(1) = EY_1$. For example, application of the theorem to $\psi_0(1) = EY_1$ and $\psi_0(0) = EY_0$, and the delta-method to $\psi_0 = \psi_0(1)/\psi_0(0) = EY_1/EY_0$, teaches us that the efficient influence curve of the causal relative risk $\psi_0 = \Psi_{q_0}(P_0)(1)/\Psi_{q_0}(P_0)(0)$ (indexed by q_0) can be represented as

$$D_{RR,q_0} = \frac{1}{\psi_0(0)} D_{1,q_0} - \frac{\psi_0(1)}{\psi_0(0)^2} D_{0,q_0},$$

where D_{1,q_0} denotes the efficient influence curve of $\psi_0(1) = EY_1$ and D_{0,q_0} denotes the efficient influence curve of $\psi_0(0) = EY_0$. Using that each of the two efficient influence curves D_{0,q_0}, D_{1,q_0} are case control weighted versions of the efficient influence curves D_0^* and D_1^* in model $\mathcal{M}^*$, it follows that we can also represent the efficient influence curve D_{RR,q_0} as

$$\begin{aligned}D_{RR,q_0} &= \frac{1}{\psi_0(0)}\{q_0 D_1^*(W_1, A_1, 1) + (1 - q_0)D_1^*(W_2, A_2, 0)\} \\ &\quad - \frac{\psi_0(1)}{\psi_0(0)^2}\{q_0 D_0^*(W_1, A_1, 1) + (1 - q_0)D_0^*(W_2, A_2, 0)\} \\ &= q_0 D_{RR}^*(W_1, A_1, 1) + (1 - q_0)D_{RR}^*(W_2, A_2, 0),\end{aligned}$$

where

$$D_{RR}^*(W, A, Y) = \frac{1}{\psi_0(0)} D_1^*(W, A, Y) - \frac{\psi_0(1)}{\psi_0(0)^2} D_0^*(W, A, Y).$$

We note that D^*_{RR} is the efficient influence curve of $\psi_0 = EY_1/EY_0$ in the nonparametric model $\mathcal{M}^*$, not treating the parameter ψ_0 as being indexed by q_0.

5.2 Variance of efficient influence curve of causal relative risk at $q_0 \approx 0$.

From the representation of the efficient influence curve of the relative risk, it follows that if q_0 and thereby $\psi_0(0)$ is very small, then the variance of the efficient influence curve D_{RR} can be very large if $D^*_0(W_2, A_2, 0)$ is a non constant random variable with a variance which is not proportional to $1/q_0^2$ or $1/\psi_0(0)^2$. As a consequence, it is fundamental that Q^*_0 is itself proportional to q_0, which is a reasonable assumption, and, for the sake of estimation, the estimators Q^*_n need to be proportional to q_0 as well.

We have the following result formalizing this.

Theorem 11 *We refer to the definition D^*_a being the efficient influence curve of $\psi^*_0(a) = E^*_0 Y_a$, $a \in \{0,1\}$ provided in previous subsection. Let*

$$\begin{aligned} D_{aq_0}(Q^*_0, g^*_0, \psi_0(a))(O) \equiv\ & q_0 D^*_a(Q^*_0, g^*_0, \psi_0(a))(W_1, A_1, 1) \\ & +(1-q_0)\frac{1}{J}\sum_{j=1}^{J} D^*_a(Q^*_0, g^*_0, \psi_0(a))(W_2^j, A_2^j, 0). \end{aligned}$$

We have that $D_{aq_0}(Q^*_0, g^*_0, \psi_0(a))$ equals the efficient influence curve of the parameter $\psi_{q_0}(a)$ and the model implied by the nonparametric model $\mathcal{M}^*$ and knowing q_0.

We have that D_{aq_0} is double robust:

$$E_0 D_{aq_0}(Q^*, g^*, \psi_0(a)) = 0 \text{ if either } g^* = g^*_0 \text{ or } Q^* = Q^*_0,$$

and in both cases we need that $g^*(1 \mid W) > 0$ a.e.

The variance of $D_{aq_0}(Q^*, g^*_0, \psi_0(a))$ under P_0 is $O(q_0)$ if

$$Q^* = q_0 \frac{1}{1-q_0} \frac{\tilde{Q}/(1-\tilde{Q})}{1 + c_0 \tilde{Q}/(1-\tilde{Q})},$$

or equivalently $Q^*/(1-Q^*) = q_0/(1-q_0)\tilde{Q}/(1-\tilde{Q})$, and $\tilde{Q}/(1-\tilde{Q})$ is bounded.

Thus, in order to construct estimators of $\psi_0(a)$ that have standard error proportional to q_0 it is crucial that we restrict our estimators of Q_0^* to estimators of the form $q_0 Q_n^\#$ for nicely bounded $Q_n^\#$ or, that it can be represented as a logistic regression with an intercept $\log c_0$ and bounded function in the covariates. In that manner, our resulting double robust locally efficient estimators of the causal relative risk or odds ratio will have a bounded influence curve at $q_0 \approx 0$.

5.3 Double robust locally efficient estimator of treatment specific mean and causal relative risk/odds ratio for case control design I.

So we can construct a double robust locally efficient estimator of $\psi_0(a)$ as follows. Let $\tilde{Q}_n$ be an estimator based on fitting a logistic regression model for Q_0^* ignoring the case control sampling: i.e., for some working model $\mathcal{Q}^*$ for $Q_0^*(A, W)$ let

$$\tilde{Q}_n = \arg\max_{Q^* \in \mathcal{Q}^*} \sum_{i=1}^n \log Q^*(W_{1i}, A_{1i}) + \frac{1}{J}\sum_{j=1}^J \log(1 - Q^*(W_{2i}^j, A_{2i}^j)).$$

We can now map this into an estimator Q_{n,q_0}^* of Q_0^*:

$$\begin{aligned} Q_{n,q_0}^* &= \frac{c_0\tilde{Q}_n/(1-\tilde{Q}_n)}{1 + c_0\tilde{Q}_n/(1-\tilde{Q}_n)} \\ &= q_0 \frac{1}{1-q_0} \frac{\tilde{Q}_n/(1-\tilde{Q}_n)}{1 + c_0\tilde{Q}_n/(1-\tilde{Q}_n)} \\ &\equiv q_0 Q_{n,q_0}^\#, \end{aligned}$$

Equivalently, we can add the intercept $\log c(q_0)$ to the log odds of the fit $\tilde{Q}_n$.

Let g_n^* be an estimator of g_0^*. For example, given a working model $\mathcal{G}^*$ for g_0^*, let

$$g_n^* = \arg\max_{g^* \in \mathcal{G}^*} \sum_{i=1}^n q_0 g^*(A_{1i} \mid W_{1i}) + (1-q_0) g^*(A_{2i} \mid W_{2i}).$$

Now, define $\psi_n(a)$ as the solution of the efficient influence curve estimating equation $P_n D_{aq_0}(Q_{n,q_0}^*, g_n^*, \psi(a)) = 0$ given by

$$\psi_n(a) = \frac{1}{n}\sum_{i=1}^n D_{aq_0}(Q_{n,q_0}^*, g_n^*)(O_i)$$

$$= q_0 \frac{1}{n} \sum_{i=1}^{n} Q^*_{n,q_0}(W_{1i}, a) \left(1 - \frac{I(A_{1i} = a)}{g^*_n(a \mid W_{1i})}\right) + q_0 \frac{I(A_{1i} = a)}{g^*_n(a \mid W_{1i})}$$
$$+(1 - q_0) \frac{1}{J} \sum_j Q^*_{n,q_0}(W^j_{2i}, a) \left(1 - \frac{I(A^j_{2i} = a)}{g^*_n(a \mid W^j_{2i})}\right).$$

Note that for $q_0 \approx 0$, this estimator is well approximated (also for purpose of causal relative risk of odds ratio) by

$$\psi_n(a) \approx q_0 \frac{1}{n} \sum_{i=1}^{n} \frac{I(A_{1i} = a)}{g^*_n(a \mid W_{1i})} + \frac{1}{J} \sum_j Q^{\#}_{n,q_0}(W^j_{2i}, a) \left(1 - \frac{I(A^j_{2i} = a)}{g^*_n(a \mid W^j_{2i})}\right).$$

5.4 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

$$\epsilon_n = \arg\max_{\epsilon} \sum_i q_0 \log Q^*(W_{1i}, A_{1i}) + (1 - q_0) \frac{1}{J} \sum_j \log(1 - Q^*(W^j_{2i}, A^j_{2i}))$$

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 $\sqrt{n}$).

6 Estimation of semi-parametric logistic regression models based on case-control sampling.

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 unspecified function r_0. We refer to this model $\{Q_{\beta,r} : \beta, r\}$ for Q_0^* as a semi-parametric logistic regression model.

We first 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^* and after that we consider the corresponding case-control weighted targeted MLE for case-control sampling from P_0^*.

Firstly, we are concerned with construction of the nuisance tangent space of the unspecified r_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 $P^*_{0\epsilon}(Y = 1 \mid A, W) = \frac{1}{1+\exp(-\{A\beta_0 W + r_0(W) + \epsilon h(W)\})}$ for arbitrary functions h. This results in the nuisance tangent space

$$T_{nuis,r_0}(P^*_0) = \{h(W)(Y - Q^*_0(A, W)) : h\}.$$

In order to find the efficient score we wish to construct a path $Q^*_0(\epsilon)$ through Q^*_0 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

$$\begin{aligned} 0 &= E^*_0(AW + h_1(W))(Y - Q^*_0(A, W))h(W)(Y - Q^*_0(A, W)) \\ &= E^*_0(AW + h_1(W))h(W)\sigma^2_0(A, W)), \end{aligned}$$

where $\sigma^2_0(A, W) = Q^*_0(A, W)(1 - Q^*_0(A, W))$. 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 the conditional expectation is w.r.t. the conditional distribution g^*_0 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 based on a sample of P_0^*. Let Q^{*0}, g^{*0} be initial estimators of Q_0^*, g_0^*, where Q^{*0} is defined by (β^0, r^0). Construct the path $Q^{*0}_{h_1^*(Q^{*0},g^{*0})}(\epsilon)$ and compute the MLE ϵ_n^0 of ϵ. This corresponds with fitting a logistic regression model in covariate $AW + h_1^*(Q^{*0}, g^{*0})$, 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 updating 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 can now be derived from the fact that it solves this estimating equation and statistical inference proceeds accordingly.

Let's now consider a case-control sample. We now set our initial estimate Q^{*0} above equal to Q^*_{n,q_0} obtained by adding an intercept $\log c_0$ into a weighted logistic regression fit $\tilde{Q}_n$ in which the cases get weight 1 and the controls receive a weight $\bar{q}_0(M_1)$. Secondly, in each estimation step of the iterative targeted MLE we assign weights q_0 and $\bar{q}_0(M_1)$ to the cases and controls, respectively. The resulting case-control weighted targeted MLE β_n^k, r_n^k now solves

$$0 = \sum_{i=1}^{n} q_0 D^*(\beta_n^k, r_n^k, g_n^{*0})(W_{i1}, A_{1i}, 1) + \frac{\bar{q}_0(M_{1i})}{J} \sum_{j=1}^{J} D^*(\beta_n^k, r_n^k, g_n^{*0})(W_{2i}^j, A_{2i}^j, 0)$$

up till a user supplied numerical precision. Statistical inference for this estimator β_n can now be based on the fact that it solves this estimating equation, or one could apply the bootstrap.

7 Targeted MLE of realistic marginal structural models for case-control studies.

The data structure on each experimental unit is $O^* = (W, A, Y) \sim P_0^*$, where W is a collection of baseline covariates, A is a treatment variable, and Y is an outcome of interest. Given a case control sample of n i.i.d. copies $O_1, \ldots, O_n$, 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 of realistic treatment options closest to a, where the latter "realistic" and "closest" need to be defined appropriately. We refer to such rules avoiding treatment assignments which are not supported by the treatment mechanism g^*_0 as realistic treatment rules. We refer to van der Laan and Petersen (2007) for a general class of causal models for realistic treatment rules.

We consider a model in which the full data distribution P^*_{X0} is unspecified. 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:

$$P^*_0(A = d(a)(W) \mid W) > \delta \text{ almost everywhere, for all } a \in \mathcal{A}_1. \tag{10}$$

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:

$$g_0^*(a \mid X) = P_0^*(A = a \mid X) = P_0^*(A = a \mid W).$$

The assumption (10) 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.

Typical models are models $m(a, v \mid \beta) = \beta(a, V)$, $m(a, v \mid \beta) = \exp \beta(a, v))$, and $m(a, v \mid \beta) = 1/(1 + \exp(-\beta(1, a, v))$ for additive effects, multiplicative effects, and odds-ratio effects, respectively.

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. Our goal is to construct a targeted MLE of β_0 based on a case-control sample.
Important identity: Under a mild regularity condition, it follows that $\beta_0 = \beta(Q_{01}^*, Q_{02}^*)$ solves

$$\begin{aligned} 0 &= E_{0V}^* \sum_{a \in \mathcal{A}_1} h(a, V) \frac{d}{d\beta_0} m(a, V \mid \beta_0)(E_0^*(Y(d(a)) \mid V) - m(a, V \mid \beta_0)) \\ &= E_{Q_{01}^*} \sum_{a \in \mathcal{A}_1} h(a, V) \frac{d}{d\beta_0} m(a, V \mid \beta_0)(E_0^*(Y(d(a)) \mid W) - m(a, V \mid \beta_0)) \end{aligned}$$

$$= \quad E_{Q_{01}^*} \sum_{a \in \mathcal{A}_1} h(a, V) \frac{d}{d\beta_0} m(a, V \mid \beta_0)(Q_{02}^*(d(a), W) - m(a, V \mid \beta_0)),$$

where we defined $Q_{02}^*(d(a), W) = E_0^*(Y \mid A = d(a)(W), W)$ and Q_{01}^* is the marginal distribution of 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^* factorizes as:

$$dP^*_{Q_0^*, g_0^*}(O^*) = Q_{01}^*(W) Q_{02}^*(Y \mid A, W) g_0^*(A \mid W),$$

where the conditional density of Y, given $A = a$, W, $Q_{20}^*(\cdot \mid a, W)$, equals the conditional density of $Y(a)$, given W, 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,

$$\psi_0(a, v) = E_0^*\{E_0^*(Y \mid A = d(a)(W), W) \mid V = v\}.$$

Case-control weighted maximum likelihood estimation: Consider a logistic regression 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 case control weighted maximum likelihood estimator θ_n:

$$\theta_n = \arg\max_\theta \sum_{i=1}^n q_0 \log Q_{2\theta}^*(1 \mid A_{1i}, W_{1i}) + \bar{q}_0(M_{1i}) \frac{1}{J} \sum_j \log Q_{2\theta}^*(0 \mid A_{2i}^j, W_{2i}^j).$$

Alternatively, we use the estimator Q_{n,q_0}^* which adds the intercept $\log c_0$ into a logistic regression fit $\tilde{Q}_n$ only using the weights $\bar{q}_0(M_{1i})/J$ for the controls, while using weight 1 for the cases.

We will leave the marginal distribution of W unspecified, so that this is estimated with the case control weighted empirical probability distribution Q_{1n}^*. 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, such as the case control weighted maximum likelihood estimator according to a model $\mathcal{G}^*$ for g_0^*.

We wish to compute the case-control weighted targeted MLE for the nonparametric model targeting β_0, based on the case-control weighted initial maximum likelihood estimator $Q_{\theta_n}^*$ 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

$$\begin{aligned} D^*(Q_0^*, g_0^*)(O^*) &= \sum_{a \in \mathcal{A}_1} I(A = d(a)(W)) \frac{h(a,V) \frac{d}{d\beta_0} m(a, V \mid \beta_0)}{g_0^*(A \mid X)} (Y - Q_{02}^*(A, W)) \\ &\quad + \sum_{a \in \mathcal{A}_1} h(a,V) \frac{d}{d\beta_0} m(a, V \mid \beta_0)(Q_{02}^*(d(a), W) - m(a, V \mid \beta_0)) \\ &\equiv D_1^*(Q_0^*, g_0^*)(W, A, Y) + D_2^*(Q_0^*)(W), \end{aligned}$$

where we defined $Q_{02}^*(d(a), W) = E_{Q_0^*}(Y \mid A = d(a)(W), W)$ and $Q_{02}^*(a, W) = E_0^*(Y \mid A = a, W)$, 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^*.

This insight allows us to define the normalizing matrix as

$$\begin{aligned} &c(P_{Q_0^*, g_0^*}^*, g_0^*, \beta_0) = \\ &P_{Q_0^*, g_0^*}^* \textstyle\sum_{a \in \mathcal{A}_1} I(A = d(a)(W)) \frac{h(a,V)}{g_0^*(A|X)} \frac{d}{d\beta_0} m(a, V \mid \beta_0) \frac{d}{d\beta_0} m(a, V \mid \beta_0)^\top \\ &= E_{Q_0^*} \textstyle\sum_{a \in \mathcal{A}_1} h(a,V) \frac{d}{d\beta_0} m(a, V \mid \beta_0) \frac{d}{d\beta_0} m(a, V \mid \beta_0)^\top. \end{aligned}$$

The efficient influence curve for β_0 at P_0^* in the nonparametric model $\mathcal{M}^*$ 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 in model $\mathcal{M}^*$ can be derived (as a linear mapping applied to the vector efficient influence curve D^*) based on the δ-method. The targeted MLE presented below 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 vector 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). If $Q_{2\theta}^*$ is a logistic regression of a binary Y on A, W, then we simply

add $\epsilon C(A,W)$, where

$$C(A,W) \equiv \sum_{a \in \mathcal{A}_1} \frac{h(a,V)\frac{d}{d\beta_0} m(a,V \mid \beta_0)}{g_0^*(A \mid X)}$$

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 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. If $m(a,V \mid \beta)$ is a linear model, then we can choose $h(a,V) = h_1(a,V)$ with (e.g.) $h_1(a,V) = g_0^*(a \mid V)$.

The one-step targeted MLE: Given an estimate g_n^* of the treatment mechanism g_0^*, let ϵ_n be the case-control weighted maximum likelihood estimator

$$\epsilon_n = \arg\max_{\epsilon} \sum_{i=1}^{n} \log q_0 Q_{2\theta_n}^*(\epsilon)(1 \mid W_{1i}, A_{1i}) + \frac{\bar{q}_0(M_{1i})}{J} \sum_{j=1}^{J} \log Q_{2\theta_n}^*(\epsilon)(0 \mid W_{2i}^j, A_{2i}^j).$$

We call $\beta_n = \beta(Q_{1n}^*, Q_{2\theta_n}^*(\epsilon_n))$ corresponding with the updated $Q_{\theta_n}^*(\epsilon_n)$ the one step targeted MLE of β_0 and if $C(a,W)$ does not depend on Q^*, then this is also the iterative targeted MLE (since the next update gives an $\epsilon_n = 0$ and thereby does not change the estimate). Either way, we let β_n be the final update after convergence of the iterative updating process has been achieved, which, for ϵ-extensions of the type presented above, occurs in a single step.

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

$$0 = \frac{1}{n}\sum_{i=1}^{n} q_0 D_2^*(\beta_n, Q^*_{\theta_n}(\epsilon_n))(W_{1i}) + \frac{\bar{q}_0(M_{1i})}{J}\sum_j D_2^*(\beta_n, Q^*_{\theta_n}(\epsilon_n))(W^j_{2i}),$$

where

$$D_2^*(\beta, Q^*)(W) = \sum_{a\in\mathcal{A}_1} h(a,V)\frac{d}{d\beta}m(a,V \mid \beta)(m(a,V \mid \beta) - Q_2^*(d(a),W))).$$

Equivalently, one can view β_n as a case-control weighted least squares solution of the regression of $Q^*_{2\theta_n}(\epsilon_n)(d(a), W_i)$ on the realistic MSM $m(a, V_i \mid \beta)$:

$$\beta_n = \arg\min_\beta \sum_{a\in\mathcal{A}_1} P_{n,q_0} h(a,V)(Q^*_{2\theta_n}(\epsilon_n)(d(a),W) - m(a,V \mid \beta))^2,$$

where P_{n,q_0} denotes the case-control weighted empirical distribution putting mass q_0/n on (W_{1i}, A_{1i}) and $\bar{q}_0(M_{1i})/(nJ)$ on (W^j_{2i}, A^j_{2i}), $j = 1, \ldots, J$. Recall that for a function $D^*(O^*)$ and corresponding $D_{q_0}(O)$ we have $P_{n,q_0}D^* = P_n D_{q_0}$.

The targeted MLE as double robust estimating function based estimator: For the purpose of statistical inference it is also helpful to note that

$$\begin{array}{l} 0 = \sum_{i=1}^n q_0 D_1^*(Q^*_{\theta_n}(\epsilon_n), g_n^*)(W_{1i}, A_{1i}, 1) + \\ \sum_{i=1}^n \bar{q}_0(M_{1i})\frac{1}{J}\sum_{j=1}^J D_1^*(Q^*_{\theta_n}(\epsilon_n), g_n^*)(W^j_{2i}, A^j_{2i}), \end{array}$$

so that we also have

$$0 = \sum_i q_0 D^*(Q^*_{\theta_n}(\epsilon_n), g_n^*)(W_{1i}, A_{1i}, 1) + \frac{\bar{q}_0(M_{1i})}{J}\sum_j D^*(Q^*_{\theta_n}(\epsilon_n), g_n^*)(W^j_{2i}, A^j_{2i}, 0)).$$

Let's now use the estimating function representation of the efficient influence curve in model $\mathcal{M}^*$,

$$\begin{aligned} D^*(\beta, Q^*, g^*) &= \sum_{a\in\mathcal{A}_1} I(A = d(a)(W))\frac{h(a,V)\frac{d}{d\beta}m(a,V \mid \beta)}{g^*(A \mid X)}(Y - Q_2^*(A,W)) \\ &\quad + \sum_{a\in\mathcal{A}_1} h(a,V)\frac{d}{d\beta}m(a,V \mid \beta)(Q_2^*(d(a),W) - m(a,V \mid \beta)), \end{aligned}$$

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 $Q_n^* = Q_{\theta_n}^*(\epsilon_n)$ solves the case-control weighted efficient influence curve equation, $P_n D_{q_0}(Q_{\theta_n}^*(\epsilon_n), g_n^*) = 0$ implies that β_n solves the case-control weighted $P_n D_{q_0}(\beta_n, Q_{\theta_n}^*(\epsilon_n), g_n^*) = 0$. Thus the targeted MLE β_n is a solution of the double robust IPTW estimating function:

$$0 = \sum_i D_{q_0}(\beta_n, Q_{\theta_n}^*(\epsilon_n), g_n^*).$$

As a consequence, we can analyze β_n in the same manner as we analyze the double robust IPTW estimator β_{nDR} solving $0 = \sum_i D_{q_0}(\beta, Q_n^*, g_n^*)$ 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 and Robins (2002)).

Statistical Inference: Thus (see van der Laan and Robins (2002)), 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 $c_0^{-1} D_{q_0}(\beta_0, Q^*, g_0^*)$, 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):

$$\beta_n - \beta_0 = \frac{1}{n} \sum_{i=1}^n c_0^{-1} D_{q_0}(\beta_0, Q^*, g_0^*)(O_i) + o_P(1/\sqrt{n}).$$

We suggest that this influence curve, as in prospective sampling, can also be used for conservative inference in the case that g_0^* is estimated according to a model, though this will need to be formally verified. 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.

Variance of estimator of $P_0^*(Y_a = 1 \mid V)$ at $q_0 \approx 0$: Let $D^*(Q_0^*, g_0^*)(O^*)$ be the non-standardized efficient influence curve in model $\mathcal{M}^*$ presented above, thus not multiplied with c_0^{-1}. Let $D_{q_0}(Q_0^*, g_0^*)(O)$ be the corresponding case-control weighted function. We note that if $Q_{02}^* = q_0 Q_{02}^{\#}$ for some bounded $Q_{02}^{\#}$, then we have that $D_{q_0}(Q_0^*, g_0^*)$ can be represented as q_0 times a bounded function. As a consequence, the variance of $D_{q_0}(Q^*, g_0^*)$ for such Q^* is proportional to q_0^2, so that the variance of the targeted MLE of $m(a, v \mid \beta_0)$ behaves as $O(q_0^2/n)$. As a consequence, we can robustly estimate causal relative risk and odds ratio effects at $q_0 \approx 0$.

8 Estimation of causal parameters for case control design I nested in a randomized trial with unknown incidence probability.

In this section we address estimation of causal parameters when one does not know q_0, but one knows the treatment mechanism. The estimators imply immediate analogues for observational case-control studies in which $q_0 \approx 0$, by simply estimating the treatment mechanism based on the control observations only.

8.1 Randomized trial example.

In previous sections, in the case that q_0 is known, we presented double robust locally efficient estimators of causal parameters. In the special case that g_0^* is known these estimators are guaranteed to always be consistent and asymptotically linear.

The IPTW estimators presented in this section are appropriate in the case that q_0 is unknown and g_0^* is known, in which case the proposed IPTW estimators of the causal relative risk and odds ratio are also always consistent and asymptotically linear. In this first subsection we discuss a randomized trial application.

Consider a randomized trial in which one samples i.i.d. $O_i^* = (W_i, A_i, Y_i = Y_i(A_i))$, $i = 1, \ldots, N$, $A_i \in \{0, 1\}$ is binary, and $P_0^*(A_i = 1 \mid X_i) = 0.5$ (say). Suppose now that at baseline one has taken a tissue sample from each patient which can be utilized to measure various markers of interest. After having run the randomized trial one wishes to design a follow up study in which one determines markers which are strong effect modifiers of the effect of treatment. Let's denote these J markers with V_j, $j = 1, \ldots, J$. For the sake of illustration suppose that the parameters targeted in such a follow up study are defined by the linear marginal structural model $E(Y(a) \mid V_j) = \beta_0 + \beta_1 a + \beta_2 V_j + \beta_3 a V_j$, where β_3 defines the parameter representing treatment effect modification of V_j. Such a study can be useful to determine future phase III trials of interest, since biomarkers which are strong effect modifiers might define sub-populations in which the treatment is particularly effective. Similarly, one might wish to assume a logistic marginal structural model.

Suppose now that it is actually very expensive to do this testing so that

one does not wish to run the biomarker assays on each patient, but on a relatively small set of patients. However, the proportion of patients with $Y_i = 1$ among the N patients is small so that selecting a random sample of the N patients for the biomarker follow up study would be a very ineffective study.

Therefore, one might run a case-control study by randomly sampling a case, and for each case one samples J controls, and one repeats this experiment n times to obtain a new sample consisting of n cases and nk controls. This data set can now be represented as $(V_{1i}, W_{1i}, A_{1i}) \sim P(V, W, A \mid Y = 1)$, $(V_{0i}^j, W_{2i}^j, A_{2i}^j) \sim P(V, W, A \mid Y = 0)$, $j = 1, \ldots, J$, $i = 1, \ldots, n$.

It should be remarked that an application of the IPTW-estimators presented in this section would ignore the controls in the N i.i.d observations O_i^* which are not sampled, while an efficient analysis would also use these controls to improve efficiency (Molinaro et al. (2005)).

8.2 Marginal structural logistic regression models for the causal odds ratio.

In this subsection we address estimation of the unknown parametes in a marginal structural logistic regression model modelling the causal effect of treatment on the odds-ratio scale. We have the following formal result.

Theorem 12 *Consider a marginal structural logistic regression model*

$$E(Y_a \mid V) = m(a, V \mid \beta_0) = \frac{1}{1 + \exp(-\beta_0 C(a, V))},$$

for a vector-valued $C(a, V)$ with $C(a, V)(0) = 1$ so that $\beta_0(0)$ denotes the intercept. Let $O^* = (W, A, Y)$ and let P_0^* be its distribution, and let $g_0^*(a \mid w) = P_0^*(A = a \mid W = w)$ be known.

Let $O = ((W_1, A_1), (W_2, A_2))$ be the experimental unit generated by case-control design I, where, for simplicity, we consider the case that we have one control for each case (i.e., $J = 1$).

Consider the following IPTW-estimating functions of O^* for β_0 indexed by h:

$$D_h^*(\beta_0)(O^*) = \frac{h(A, V)}{g_0^*(A \mid W)}(Y - m(A, V \mid \beta_0)).$$

We note that $P_0^* D_h(\beta_0) = 0$ for all h, under the assumption that $\sup_a h(a, V)/g_0^*(a \mid W)$ is bounded a.e.

Consider the following class of corresponding IPTW-estimating functions of O for β_0

$$D_h(\beta)(O) \;=\; D_h^*(\beta)(W_1, A_1, 1) + D_h^*(\beta)(W_2, A_2, 0).$$

If $P_0 D_h^*(\beta_0) = 0$, then there exists a β_0' with $\beta_0'(1, \ldots, d) = \beta_0(1, \ldots, d)$ satisfying

$$P_0 D_h(\beta_0') = 0,$$

As a consequence, for case-control design I, one can estimate the non-intercept coefficients in the marginal structural logistic regression model with weighted maximum likelihood estimation for the logistic regression of Y on A, V according to the MSM model, using weights $g_0^*(A_1 \mid V_1)/g_0^*(A_1 \mid W_1)$ for the cases and $g_0^*(A_2 \mid V_2)/g_0^*(A_2 \mid W_2)$ for the controls and further ignoring the case-control sampling. That is, we can estimate β_0 with the solution of

$$0 = \sum_{i=1}^{n} D_h(\beta_n)(O_i),$$

or equivalently, we can set

$$\beta_n = \arg\max_\beta \textstyle\sum_i \frac{g_0^*(A_{1i}|V_{1i})}{g_0^*(A_{1i}|W_{1i})} \log m_\beta(V_{1i}, A_{1i}) + \frac{g_0^*(A_{2i}|V_{2i})}{g_0^*(A_{2i}|W_{2i})} \log(1 - m_\beta(V_{2i}, A_{2i})).$$

In other words, one can fit the (odds-ratio part) of the marginal structural logistic regression model with the IPTW estimator ignoring the case-control sampling.

We note that this result naturally generalizes to data structures O^* including a time-dependent treatment and marginal structural logistic regression models modelling the causal effect of the multiple time point treatment on the binary outcome.

This result is related and based on the same principles as the remark on case-control studies in Robins (1999) in the context of direct effect estimation and this IPTW-estimator is practically investigated in Mansson et al. (2007).

For the sake of completeness, we will now prove this result.

Proof of Theorem 12: Firstly, we note that replacing $g_0^*(A \mid W)$ by a $g_0^*(A \mid V)$ in P_0^* corresponds with assuming that A is independent of (Y_0, Y_1), given V, and we denote the corresponding manipulated version of P_0^* with P_0^{*m}. We first define a correct estimation procedure under case-control sampling from this manipulated population distribution P_0^{*m}.

In that case, $E_0^{*m}(Y_a \mid V) = E_0^{*m}(Y \mid A = a, V)$. Since a standard maximum likelihood estimator of the logistic regression model for $E(Y \mid A = a, V)$ correctly estimates the odds-ratio part of $E(Y \mid A, V)$ it follows that the marginal structural model can be fitted with standard maximum likelihood estimation, and that the resulting estimator is consistent for the non-intercept components of β_0. This maximum likelihood estimator solves the estimating equation corresponding with the estimating function of the form $D_h^m(\beta_0) \equiv h(A, V)(Y - m(A, V \mid \beta_0))$, where $h = d/d\beta m/(m(1-m))$. Thus, it follows that for each h $P_0^{*m} D_h^m(\beta_0') = 0$ for a β_0' which agrees with the true β_0 up till its intercept, and for this same β_0' we have

$$
\begin{aligned}
&0 = P_0^m h(A_1, V_1)(1 - m(A_1, V_1 \mid \beta_0')) + h(A_2, V_2)(0 - m(A_2, V_2 \mid \beta_0')) \\
&= q_0 \textstyle\sum h(a, v)(1 - m(a, v \mid \beta_0'))dP_0^{*m}(w, a, 1) \\
&+(1 - q_0) \textstyle\sum h(a, v)(0 - m(a, v \mid \beta_0'))dP_0^{*m}(w, a, 0).
\end{aligned}
$$

In other words, the empirical summation $\sum_i D_h^m(\beta_0')(W_{1i}, A_{1i}, 1) + D_h^m(\beta_0')(W_{2i}, A_{2i}, 0)$ ignoring the case-control sampling has mean zero at this β_0' under the manipulated case-control samping distribution P_0^m.

This basic latter identity will now be used to show that the Inverse Probability of Treatment Weighted (IPTW) estimating function $D_h(\beta_0')$ is unbiased under P_0. Note that

$$
\begin{aligned}
&P_0 \tfrac{g_0^*(A_1|V_1)}{g_0^*(A_1|W_1)} h(A_1, V_1)(1 - m(A_1, V_1 \mid \beta_0')) \\
&+P_0 \tfrac{g_0^*(A_2|V_2)}{g_0^*(A_2|W_2)} h(A_2, V_2)(0 - m(A_2, V_2 \mid \beta_0')) = \\
&q_0 \textstyle\int \frac{g_0^*(a|v)}{g_0^*(a|w)} h(a, v)(1 - m(a, v \mid \beta_0'))dP_0^*(w, a, 1) \\
&+(1 - q_0) \textstyle\int \frac{g_0^*(a|v)}{g_0^*(a|w)} h(a, v)(0 - m(a, v \mid \beta_0')dP_0^*(w, a, 0) \\
&= q_0 \textstyle\int h(a, v)(1 - m(a, v \mid \beta_0'))dP_0^{*m}(w, a, 1) \\
&+(1 - q_0) \textstyle\int h(a, v)(0 - m(a, v \mid \beta_0'))dP_0^{*m}(w, a, 0) \\
&= 0.
\end{aligned}
$$

by our previously established identity.

The more general result also follows by noting that the above result applies to each h. This completes the proof. □

8.3 Case-only IPTW estimators for the causal relative risk.

In this subsection we consider a simple estimator of a causal relative risk in a nonparametric model.

We start out with proving the following theorem which is the basis of the IPTW-estimator only using the cases, as presented below.

Theorem 13 *Suppose $D^*(P_0^*) = D^*(\psi_0, \eta_0^*)$ is an estimating function for ψ_0 with nuisance parameter η_0^*. Assume that*

$$D^*(\psi, \eta) = D_1^*(\eta) - \psi$$

for some D_1^*. Then ψ_0 satisfying $P_0 D_{q_0}(\psi_0, \eta_0) = 0$, with $D_{q_0}(O) = q_0 D^*(W_1, A_1, 1) + \frac{1-q_0}{J} \sum_j D^*(W_0^j, A_0^j, 0)$, is given by

$$\psi_0 = P_0 D_1^*(W_2, A_2, 0) + q_0 P_0 \{D_1^*(W_1, A_1, 1) - \frac{1}{J} \sum_j D_1^*(W_2^j, A_2^j, 0)\}.$$

In particular, if $D_1^*(W, A, 0) = 0$ a.e., then

$$\psi_0 = q_0 P_0 D_1^*(W_1, A_1, 1).$$

One can apply this theorem to EY_1 and EY_0 for D^* satisfying $D_1^*(W, A, 0) = 0$. Consider $\psi_0(1) = EY_1 = EE_{P_0^*}(Y \mid A = 1, W)$ and consider the estimating function $D^*(P_0^*)(W, A, Y) = YA/g_0^*(1 \mid W) - \psi_0(1)$. Then $D(\psi_0(1), g_0^*)(W, A, 0) = -\psi_0(1)$ so that it follows that

$$\psi_0(1) = q_0 P_0 \frac{A_1}{g_0^*(1 \mid W_1)}.$$

This yields the following identifiability result:

$$\frac{\psi_0(1)}{q_0} = \frac{EY(1)}{q_0} = P_0 \frac{A_1}{g_0^*(1 \mid W_1)}.$$

Note that this parameter represents a relative risk measure representing an effect of an intervention relative to the current population proportion and note that the identifiability result does not require knowing q_0. We refer to the resulting estimator as the case-only IPTW estimator.

Similarly,

$$\psi_0(0) = q_0 P_0 \frac{1 - A_1}{g_0^*(0 \mid W_1)},$$

giving us the identifiability result

$$\psi_0(0)/q_0 = P_0(1 - A_1)/g_0^*(0 \mid W_1).$$

Thus, if g_0^* is known, then one can identify the following causal parameters of interest

$$\begin{aligned}\frac{\psi_0(1)-\psi_0(0)}{q_0} &= \frac{E(Y(1)-Y(0))}{E^*Y}\\ \frac{\psi_0(1)}{q_0} &= \frac{EY(1)}{E^*Y}\\ \frac{\psi_0(0)}{q_0} &= \frac{EY(0)}{E^*Y}.\end{aligned}$$

Similarly, it follows that we can identify the causal relative risk $\psi_{0RR} \equiv \psi_0(1)/\psi_0(0)$:

$$\psi_{RR0} \equiv \frac{\psi_0(1)}{\psi_0(0)} = \frac{P_0A_1/g_0^*(1 \mid W_1)}{P_0(1-A_1)/g_0^*(0 \mid W_1)}.$$

One can also identify the causal odds ratio:

$$\psi_{OR0} \equiv \frac{\psi_0(1)/(1-\psi_0(1))}{\psi_0(0)/(1-\psi_0(0))} = \frac{P_0A_1/g_0^*(1 \mid W_1)/(1-P_0A_1/g_0^*(1 \mid W_1))}{P_0(1-A_1)/g_0^*(0 \mid W_1)/(1-P_0(1-A_1)/g_0^*(0 \mid W_1))},$$

but estimation of the causal odds-ratio was addressed in the previous subsection with a preferred estimation strategy.

The corresponding case only IPTW estimators of these causal parameters have robust influence curves even if $q_0 \approx 0$. For example, the influence curve of the case-only IPTW estimator

$$\psi_n = \frac{\sum_{i=1}^n I(A_{1i}=1)/g_0^*(1 \mid W_{1i})}{\sum_{i=1}^n I(A_{1i}=0)/g_0^*(0 \mid W_{1i})}$$

of the causal relative risk $\psi_0(1)/\psi_0(0) = E(Y(1))/E(Y(0))$ is given by

$$IC_1(O) \equiv \frac{q_0}{\psi_0(0)}\left\{\frac{I(A_1=1)}{g_0^*(1 \mid W_1)} - \frac{\psi_0(1)}{\psi_0(0)}\frac{I(A_1=0)}{g_0^*(0 \mid W_1)}\right\}. \tag{11}$$

Note that this influence curve remains bounded for rare diseases as long as $q_0/\psi_0(0)$ remains bounded.

We note that if $q_0 \approx 0$, one can also estimate the causal relative risk with the IPTW estimator of the saturated marginal structural logistic model for $P(Y_a = 1)$, by using that the causal odds ratio approximates the causal relative risk. We believe, if $q_0 \approx 0$, then the latter approach will result in a more precise estimator.

8.4 Inverse probability of treatment weighted estimators of linear marginal structural models for rare outcomes.

The previous theorem provided us for the case that q_0 is unknown and g_0^* known with simple inverse probability of treatment weighted identifiability results for the causal relative risk parameters. These IPTW estimators appeared to be relatively stable estimation procedures, even at small q_0. In the case that $q_0 \approx 0$, but still unknown, these IPTW-estimators could be extended by replacing g_0^* by a model based estimate based on the control observations only.

We will now extend these results to linear marginal structural models by using the following theorem.

Theorem 14 *Consider an estimating function $D^*(\psi)(W, A, Y)$ for a k-dimensional parameter ψ satisfying*

$$D^*(\psi)(W, A, Y) = Y D_1^*(A, W) + D_2^*(A, W)\psi \tag{12}$$

for some $k \times 1$ vector function D_1^* and $k \times k$ matrix function D_2^*. Let ψ_0 solve $P_0^* D^*(\psi_0) = 0$. In this case,

$$0 = P_0 \left\{ q_0 D^*(\psi)(M_1, W_1, A_1, 1) + \frac{1 - q_0}{J} \sum_{j=1}^{J} D^*(\psi)(M_1, W_2^j, A_2^j, 0) \right\}$$

implies

$$\frac{\psi_0}{q_0} = \{E_0^* D_2(W, A)\}^{-1} E_0 D_1^*(A_1, W_1),$$

where

$$E_0^* D_2(W, A) = E \left\{ q_0 D_2(W_1, A_1) + \frac{1 - q_0}{J} \sum_{j=1}^{J} D_2(W_2^j, A_2^j) \right\}.$$

The proof of Theorem 14 is straightforward and therefore omitted. This theorem teaches us that by using estimating functions D^* satisfying the mentioned structure (12), one can obtain closed estimators of ψ_0/q_0 which are stable even for small values of q_0: i.e., these estimators have bounded influence curve uniformly in $q_0 \approx 0$). In addition, one can obtain these estimators without knowing the vale of q_0 as long as one knows that $q_0 \approx 0$.

As an example, let's consider a causal effect model describing how the treatment specific mean changes as a function of treatment and an adjustment variable V:

$$E_0^*(Y(a) \mid V) = \beta_0^\top m(a, V),$$

where, for example, $m(a, V)^\top = (1, a, V, aV)$. Let β_0 denote the true vector of values.

A least squares IPTW estimating function for β based on i.i.d sampling of $O^* = (W, A, Y)$ is given by

$$D^*(\beta)(W, A, Y) = \frac{m(A, V)}{g_0^*(A \mid W)}(Y - \beta^\top m(A, V)).$$

which satisfies that $P_0^* D^*(\beta_0) = 0$ if $\sup_a m(a, V)/g_0^*(a \mid W) < \infty$ a.e. In general, we can choose

$$D^*(\beta)(W, A, Y) = \frac{h(A, V)}{g_0^*(A \mid W)}(Y - \beta^\top m(A, V)),$$

for arbitrary function h, where, for example, one can select

$$h(A, V) = \frac{m(A, V)}{\beta^\top m(A, V)(1 - \beta^\top m(A, V))},$$

which corresponds with weighted least squares. For simplicity, let's consider the case that $h(A, V) = m(A, V)$.

We have that D^* indeed satisfies the wished linear structure (12):

$$D^*(\beta)(O^*) = Y \frac{m(A, V)}{g_0^*(A \mid W)} - \frac{m(A, V)m^\top(A, V)}{g_0^*(A \mid W)}\beta.$$

Thus, an application of Theorem 14 teaches us that

$$\frac{\beta_0}{q_0} = \left\{ E_0^* \frac{m(A, V)m^\top(A, V)}{g_0^*(A \mid W)} \right\}^{-1} E_0 \frac{m(A_1, V_1)}{g_0^*(A_1 \mid W_1)},$$

where the normalizing matrix

$$c_0 = \left\{ E_0^* \frac{m(A, V)m^\top(A, V)}{g_0(A \mid W)} \right\} = E_0^* \sum_a m(a, V)m^\top(a, V)$$

is identified as

$$E_0 q_0 \sum_a m(a, V_1) m^\top(a, V_1) \\ + \frac{1-q_0}{J} \sum_{j=1}^J \sum_a m(a, V_2^j) m^\top(a, V_2^j).$$

Since the latter depends on q_0 this is not really providing the wished identifiability result. But, if $q_0 \approx 0$, then the estimator β_n can be defined as:

$$\frac{\beta_n}{q_0} = c_n^{-1} \frac{1}{n} \sum_{i=1}^n \frac{m(A_{1i}, V_{1i})}{g_0^*(A_{1i} \mid W_{1i})},$$

with

$$c_n = \frac{1}{n} \sum_{i=1}^n \frac{1}{J} \sum_{j=1}^J \sum_a m(a, V_{2i}^j) m^\top(a, V_{2i}^j),$$

which does not rely on knowing q_0.

These results can now be used to estimate a number of relative risk parameters of interest. We have $E(Y_1 \mid V) = \beta_0^\top m(1, V)$, $E(Y_0 \mid V) = \beta_0^\top m(0, V)$, so that

$$\frac{E(Y_1 \mid V) - E(Y_0 \mid V)}{E(Y_0 \mid V)} = \frac{\beta_0^\top (m(1, V) - m(0, V))}{\beta_0^\top m(0, V)} = \frac{\beta_0^\top / q_0 (m(1, V) - m(0, V))}{\beta_0^\top / q_0 m(0, V)}.$$

Thus, this conditional relative causal effect of treatment versus control is a simple function of β_0 / q_0, and can thus be estimated as above.

Another possible relative risk parameter is

$$\frac{E(Y_1 \mid V) - E(Y_0 \mid V)}{EY} = \frac{\beta_0^\top}{q_0} (m(1, V) - m(0, V)).$$

8.5 Deriving the efficient influence curve for case-control design I and unknown incidence probability

In this subsection we indicate how one would go about deriving an efficient estimator in the model in which the incidence probability is unknown, but g_0^* is known.

Consider the model $\mathcal{M}^*$ only assuming that g_0^* is known, $J = 1$, and case-control design I so that

$$dP^*_{,Q_1^*,Q^*}(O) = \frac{1}{q(Q^*, Q_1^*)(1 - q(Q^*, Q_1^*))} Q^*(W_1, A_1))(1 - Q^*(W_2, A_2)) Q_1^*(W_1) Q_1^*(W_2) \tag{13}$$

is indexed by infinite dimensional parameters $Q^*(a, w) = P^*(Y = 1 \mid A = a, W = w)$ and $Q_1^*(w) = P(W = w)$.

We stress that the following efficiency calculations apply to this particular model in which g_0^* is known, and q_0 is unknown (and thus not to the case that q_0 is known). We expect that the following can be easily generalized to general J.

We first determine the so called tangent space for this case-control design I model. We will first determine the score operator which maps the scores $h(Y \mid A, W)$ (satisfying $h(0 \mid A, W) = Q^*(A, W)/(1-Q^*(A, W))h(1 \mid A, W)$) and $h_1(W)$ (mean zero) of fluctuations $Q_\epsilon^*(Y \mid A, W)$ and $Q_{1\epsilon}^*(W)$ through $Q^*(Y \mid A, W)$ and $Q^*(W)$ into the score of $dP_{Q_\epsilon^*, Q_{1\epsilon}^*}$. This score operator is given by:

$$A^*(h, h_1) = h(1 \mid A_1, W_1) - \frac{Q^*(A_2, W_2)}{1 - Q^*(A_2, W_2)} h(1 \mid A_2, W_2) + h(W_1) + h(W_2),$$

where $h(1 \mid A_1, W_1)$ can be an arbitrary function with mean zero, and h is an arbitrary functions of W with mean zero w.r.t. $\bar{Q}(W) = Q_1(W) + Q_0(W)$, where $Q_1(w) = P(W = w \mid Y = 1)$, $Q_0(w) = P(W = w \mid Y = 0)$.

Secondly, it is well known (Bickel et al. (1993)) that the efficient influence curve of a parameter ψ_0^* of P_0^*, such as the marginal causal odds ratio in the nonparametric model, can be obtained by projecting the influence curve IC_1 of an initial regular asymptotically linear estimator of this parameter ψ_0^*, such as the IPTW estimator for the saturated logistic marginal structural model ignoring the case-control sampling as above, onto the closure of the range of the score operator A^* in the Hilbert space $L_0^2(P_0)$ endowed with inner product $\langle V_1, V_2 \rangle_{P_0} = E_{P_0} V_1(O) V_2(O)$. This insight allows us in principle to calculate the efficient influence curve, and thereby obtain a locally efficient estimator improving on the initial estimator. Since the resulting calculations are quite extensive, this is not reported here. Our intuition is that the IPTW estimator for the logistic marginal structural model is highly efficient in the case that q_0 is unknown, since not knowing q_0 makes it essentially impossible (at $q_0 \approx 0$) to estimate $E_0^*(Y \mid A, W)$ which is needed to fully exploit the covariates for efficiency gains, as in the double robust targeted MLE for the case that q_0 is known.

8.6 Discussion.

For the sake of discussion consider the case that g_0^* is known. We suggest that the mentioned IPTW-estimators for the logistic and linear marginal structural model are highly efficient in the model in which the incidence

probability is unknown. Based on that premise one needs to conclude that somehow knowing q_0, even when it is known that it is very small $q_0 \approx 0$, can truly help to gain efficiency relative to these IPTW estimators, while without knowing q_0 this gain in efficiency cannot be achieved. We believe that the crucial reason is that it requires knowing q_0 to convert a fit of a logistic regression $\tilde{Q}_n^*$ (that does not suffer from $q_0 \approx 0$ for the sake of estimation of the conditional odds-ratio), obtained with the maximum likelihood estimator ignoring the case control sampling, into a valid estimator of Q_0^* with standard error proportional to $q_0/\sqrt{n}$ (uniform in $q_0 \approx 0$). In this manner one can use covariate information, as utilized by fitting a logistic regression model correctly targeting the conditional odds ratio and then adding the intercept $\log c_0$ into the obtained logistic regression fit, to improve efficiency. Without knowing q_0, an estimator of Q_0^* obtained by using the known g_0^* will have a standard error which is too large since it will not be proportional to q_0.

8.7 Estimation of treatment mechanism.

Consider now the case that one does not know the treatment mechanism g_0^*, but that one is willing to assume a model for g_0. If q_0 would be known, then, for all these causal parameters, one would estimate g_0^* with weighted maximum likelihood estimation: given a model $\mathcal{G}$ for g_0^*

$$g_n^* = \arg\max_{g \in \mathcal{G}} \sum_{i=1}^{n} q_0 \log g(A_{1i} \mid W_{1i} + \frac{1-q_0}{J} \sum_{j=1}^{J} \log g(A_{2i}^j \mid W_{2i}^j).$$

Note that in the case that $q_0 \approx 0$, the case-control weighted maximum likelihood estimator of g_0^* would essentially only use the control observations with weight 1 and thus yield an estimate of the conditional distribution of A_2, given W_2.

Inspection of possible bias in estimator of g_0^* due to only using control observations. We note

$$\begin{aligned} P_0^*(A=1 \mid W, Y=0) &= P_0^*(A=1 \mid W) \frac{P_0^*(Y=0 \mid A=1, W)}{P_0^*(Y=0 \mid W)} \\ &\equiv P_0^*(A=1 \mid W) r_0(1,W)^{-1}. \end{aligned}$$

So we need that for q_0 small the ratio

$$r_0(1,W) = \frac{P_0^*(Y=0 \mid W)}{P_0^*(Y=0 \mid A=1, W)} \approx 1.$$

This seems a weak assumption since one also knows that

$$E_W P_0^*(Y = 0 \mid W) = 1 - q_0 \approx 1 \text{ and } E_W P_0^*(Y = 1 \mid A = 1, W) = O(q_0),$$

assuming that $g_0^*(1 \mid W)$ is bounded away from zero and 1. Thus, one expects that both numerator and denominator of $r_0(1, W)$ are close to 1 for almost all W, and thereby that $r_0 \approx 1$. Having said this, it should theoretically be possible to have certain integrals be close to 1 while another being very large, so that strictly speaking it seems that we need to make an assumption in order to guarantee that for q_0 small the required integrals behave as expected.

For each specific parameter and corresponding IPTW-estimator, we can generate more precise statements regarding this approximation of $r_0(a, W)$ in a context in which $q_0 \to 0$. For example, we wish to replace the causal relative risk by the approximation

$$\frac{E_0 \frac{I(A_1=1)}{g_0^*(1|W_1,Y=0)}}{E_0 \frac{I(A_1=0)}{g_0^*(0|W_1,Y=0)}} = \frac{E_0 \frac{I(A_1=1)}{g_0^*(1|W_1) r_0(1,W_1)}}{E_0 \frac{I(A_1=0)}{g_0^*(0|W_1) r_0(0,W_1)}}.$$

For example, if

$$r_0(a, W) = 1 + r(q_0),$$

for a constant (in W) remainder term $r(q_0)$ which converges to zero when $q_0 \to 0$, then under a weak regularity condition, it follows that

$$\frac{E_0 \frac{I(A_1=1)}{g_0^*(1|W_1,Y=0)}}{E_0 \frac{I(A_1=0)}{g_0^*(0|W_1,Y=0)}} = \frac{E_0 \frac{I(A_1=1)}{g_0^*(1|W_1)}}{E_0 \frac{I(A_1=0)}{g_0^*(0|W_1)}} + O(r(q_0)).$$

However, since $r_0(a, W)$ is averaged over W, one should only need that an integral of the remainder $r(q_0)$ as a function of W is small (e.g., $O(q_0)$). This does not seem to be a strong assumption at al given that an integral of numerator and denominator of $r(q_0)$ have to be $O(q_0)$.

To conclude, in order to have that g_n^* based on control observations only is a practically unbiased estimator of g_0 when the disease studied is very rare, we need that $r_0(a, W) = \frac{P_0^*(Y=0|W)}{P_0^*(Y=0|A=a,W)} \approx 1$ for $a \in \{0, 1\}$ for almost every W. We note that this does allow that the treatment $A = 1$ increases the risk on $Y = 1$ by a factor relative to $A = 0$. In other words, this condition requires that for most W, $P_0^*(Y = 1 \mid A = a, W)$ is small (say of the order q_0).

9 Summary, discussion and simple 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. 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.

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, 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 most applications the marginal population proportion of cases, q_0, should be known, at least within close approximation, assuming

one has made an effort to understand the target population the cases are sampled from. 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.

Allthough the main purpose of our article is the introduction, study, and application of the general methodology for analyzing case-control studies based on a known (or set value in a sensitivity analysis) incidence probability, for the sake of completeness, we also wanted to consider the case that the population proportion of cases q_0 is unknown in case-control design I. We presented various IPTW-type estimators of causal parameters relying on $q_0 \approx 0$ or that the treatment mechanism $g_0^*(A \mid W)$ is known. We also highlighted the known result (Robins (1999) and Mansson et al. (2007)) showing that one can estimate a marginal structural logistic regression model with standard IPTW-logistic regression software, again either assuming g_0^* is known or that $q_0 \approx 0$. These IPTW-estimators rely on a correctly specified model for g_0^* and require $q_0 \approx 0$. Since, without knowing q_0 and without knowing g_0^*, the IPTW-estimators target a non-identifiable parameter, we are concerned about the sensitivity of these IPTW estimators w.r.t. misspecifying g_0^*.

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 sam-

ples 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).

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 15 *(***Tangent space for case-control design I***) Consider case-*

control design I and the independence model $\mathcal{M}$ described by (7),

$$dP(P^*)(O) = P^*(W_1, A_1 \mid Y = 1) \prod_j P^*(W_2^j, A_2^j \mid Y = 0),$$

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

$$T_I(P^*) = \left\{ S^*(W_1, A_1, 1) - E^*(S^* \mid Y = 1) + \sum_j \{S^*(W_2^j, A_2^j, 0) - E^*(S^* \mid Y = 0)\} \right\},$$

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 16 *(**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

$$T^*(P^*) = \sum_\delta T_\delta^*(P^*), \tag{14}$$

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

$$T^*(P^*) = L_0^2(P_Y^*) \oplus \sum_\delta T_\delta^*(P^*), \tag{15}$$

where $L_0^2(P_Y^*)$ is the Hilbert space of functions of Y with mean zero and finite variance w.r.t. P^*. We also note that for a $S^* \in L_0^2(P^*)$, the projection of S^* on $T_\delta^*(P^*)$ is given by

$$\Pi(S^* \mid T_\delta^*(P^*)) = I(Y = \delta)\left(S^*(W, A, \delta) - E(S^* \mid Y)\right),$$

and the projection of S^* onto $T^*(P^*)$ described by the orthogonal decomposition (15) 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 17 *(***Tangent space for case-control design II***) Consider case-control design II and the independence model $\mathcal{M}$ described by (8),*

$$dP(P^*)(O) = P^*(M_1)P^*(A_1, W_1 \mid Y = 1, M_1)\prod_j P^*(A_2^j, W_2^j \mid Y = 0, M_1),$$

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

$$T_{II}(P^*) = L_0^2(M_1) \oplus \\ \left\{S^*(Z_1, 1) - E^*(S^* \mid M = M_1, Y = 1) + \textstyle\sum_j \{S^*(Z_2^j, 0) - E^*(S^* \mid M = M_1, Y = 0)\}\right\},$$

where S^* varies across $T^*(P^*)$, $Z_1 = (M_1, W_1, A_1)$ and $Z_2^j = (M_1, W_2^j, A_2^j)$.

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 18 *(***Tangent space for model $\mathcal{M}^*$ including matching variable***)*

We make the following assumption on $\mathcal{M}^*$: 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

$$T^*(P^*) = L_0^2(M) \oplus \sum_{m,\delta} T^*_{m,\delta}(P^*), \tag{16}$$

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

$$T^*(P^*) = L_0^2(P^*_M) \oplus T^*(q_0) \oplus \sum_{m,\delta} T^*_{m,\delta}(P^*), \tag{17}$$

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 (17) 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 (15). 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 (17). Given a $D^* \in T^*(P^*)$, we have*

$$\begin{aligned}
S^*_m(M,W,A,Y) &\equiv I(M=m)q_0(Y \mid m)\left\{D^*(m,W,A,Y) - E^*(D^* \mid M,Y)\right\} \\
&\in T^*(P^*). \qquad (18)
\end{aligned}$$

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

$$\begin{aligned}
&\textstyle\sum_\delta q_0(\delta \mid m)\Pi(D^* \mid T^*_{m,\delta}(P^*)) \\
&= \textstyle\sum_\delta q_0(\delta \mid m) I(M=m, Y=\delta)\left(D^*(m,W,A,\delta) - E^*(D^* \mid M,Y)\right) \\
&= I(M=m)q_0(Y \mid m)\left(D^*(m,W,A,Y) - E^*(D^* \mid M,Y)\right) \\
&= S^*_m(M,W,A,Y) \\
&\in T^*(P^*).
\end{aligned}$$

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 19 *(**Case-control weighted score is a score, Design I**)*

Consider case-control design I, its independence model $\mathcal{M}$ described by (7), and assume the tangent space $T^*(P^*)$ of $\mathcal{M}^*$ at P^* satisfies the orthogonal decomposition (15).

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

$$\begin{aligned} D_{q_0}(O) &= S^*(W_1, A_1, 1) - E^*(S^*(W, A, Y) \mid Y = 1) \\ &\quad + \sum_j \{S^*(W_2^j, A_2^j, 0) - E^*(S^*(W, A, Y) \mid Y = 0)\}. \end{aligned}$$

(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 (7), and assume the tangent space $T^*(P^*)$ of $\mathcal{M}^*$ at P^* satisfies the orthogonal decomposition (15).*

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 20 *(**Case-control weighted score is a score, Design II**)*

Consider case-control design II, its independence model $\mathcal{M}$ described by (8), and assume the tangent space $T^*(P^*)$ of $\mathcal{M}^*$ at P^* satisfies the orthogonal decomposition (17).

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 (18) 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 $D^* \in L^2_0(P^*)$,

$$q_0(1 \mid m)E^*(D^* \mid M = m, Y = 1) + q_0(0 \mid m)E(D^* \mid M = m, Y = 0) = 0.$$

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 (8), and assume the tangent space $T^*(P^*)$ of $\mathcal{M}^*$ at P^* satisfies the orthogonal decomposition (17).

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 $P(P^*)$.

Appendix: Efficient influence curve of marginal causal effect in nonparametric model for case-control design I

In this section we establish that the efficient influence curve of the marginal causal effects defined on a nonparametric model $\mathcal{M}^*$ and indexed by a fixed known q_0 for case-control design I can be represented as a case-control weighted $D_{q_0} = q_0 D^*(\cdot, 1) + (1 - q_0) D^*(\cdot, 0)$, with D^* being the efficient influence curve of the marginal causal effect defined on the nonparametric model $\mathcal{M}^*$ (and not indexed by q_0).

Our general theorems 4 and 7 teaches us that this should indeed be true but with D^* being the efficient influence curve of the marginal causal effect defined as $\Psi^*_{q_0}$ indexed by a fixed q_0, which can be shown as well (and follows from Theorem 9).

Theorem 21 *(**Efficient influence curve for case control design I**)*

Consider Case Control Design I with data structure $O = ((W_1, A_1), ((W_2^j, A_2^j) : j))$, where (W_1, A_1) has distribution $Q_1 \sim (W, A) \mid Y = 1$ and (W_2, A_2) has distribution $Q_0 \sim (W, A) \mid Y = 0$. Let $Q_1(w, a) = P(W_1 = w, A_1 = a)$ and let $Q_0(w, a) = P(W_2 = w, A_2 = a)$. Similarly, we define $Q_1(w) = P(W_1 = w)$ and $Q_0(w) = P(W_2 = w)$. We also define $Q^*(a, W) = P^*(Y = 1 \mid A = a, W)$, $Q^*_W(w) = P^*(W = w)$. Let $J = 1$.

Consider the working model that (W_1, A_1) is independent of (W_2, A_2), and no further assumptions, so that the likelihood is simply

$$p_0(O) = Q_1(W_1, A_1) Q_0(W_2, A_2),$$

and Q_1 and Q_2 are unspecified. Then, under appropriate regularity conditions, the nonparametric maximum likelihood estimator, defined by the empirical distributions of Q_{1n} and Q_{0n} of Q_1 and Q_0, of the parameter $\Psi_{q_0}(P_0)(1)$ defined by

$$\begin{aligned} P_0(Y_1 = 1) &= E_W E(Y \mid A = 1, W) \\ &= \sum_w Q^*(w, 1) Q_W^*(w) \\ &= \sum_w \frac{Q_1(W, 1) q_0}{Q_1(W, 1) q_0 + Q_0(W, 1)(1 - q_0)} \{q_0 Q_1(w) + (1 - q_0) Q_0(w)\} \\ &\equiv \Psi_{q_0}(P_0) \end{aligned}$$

is regular and asymptotically linear with influence curve

$$D_{q_0}(P_0)(O) = q_0 D^*(P_0^*)(W_1, A_1, 1) + (1 - q_0) D^*(P_0^*)(W_2, A_2, 0)$$

with $D^*(P_0^*)(W, A, Y) = (Y - Q_0^*(1, W)) I(A = 1)/g_0^*(1 \mid W) + Q_0^*(1, W) - \Psi(Q_0^*)$, $Q_0^*(a, W) = P_0^*(Y = 1 \mid A = a, W)$ and $g_0^*(a \mid W) = P_0^*(A = a \mid W)$.

This shows that in this independence model the efficient influence curve is given by $D_{q_0}(P_0)$.

Since, also under dependence of (W_1, A_1) and (W_2, A_2), this nonparametric NPMLE is a regular consistent and asymptotically linear estimator of $\Psi_{q_0}(P_0)$, it follows that $D_{q_0}(P_0)$ is also the efficient influence curve in the model in which one allows dependence between (W_1, A_1) and (W_2, A_2) (as long as they have the specified marginal distributions Q_1 and Q_0, respectively).

In general, for Case Control I, we have that the efficient influence curve for this independence or the bigger arbitrary dependence model (or any model in between) is given by

$$D_{q_0}(P_0)(O) = q_0 D^*(P_0^*)(W_1, A_1, 1) + (1 - q_0) \frac{1}{J} \sum_{j=1}^{J} D^*(P_0^*)(W_2^j, A_2^j, 0)$$

Proof. We provide the proof for $J = 1$. Let $Q_{1n}(w, a) = 1/n \sum_{i=1}^n I(W_{1i} = w, A_{1i} = a)$, $Q_{0n} = 1/n \sum_{i=1}^n I(W_{2i} = w, A_{2i} = a)$ be the empirical distributions of Q_1 and Q_0. Similarly, let $Q_{1n}^1(w) = 1/n \sum_{i=1}^n I(W_{1i} = w)$ and $Q_{0n}^1(w) = 1/n \sum_{i=1}^n I(W_{2i} = w)$ be the marginal empirical distributions of Q_1^1 and Q_0^1.

Note the NPMLE of ψ_0

$$\begin{aligned}\psi_n &= f(Q_{0n}^1, Q_{1n}^1, Q_{1n}, Q_{0n}) \\ &\equiv \sum_w \{q_0 Q_{1n}(w) + (1-q_0)Q_{0n}(w)\} \frac{Q_{1n}(w,1)q_0}{Q_{1n}(w,1)q_0 + Q_{0n}(w,1)(1-q_0)}.\end{aligned}$$

We also have $\psi_0 = f(Q_0^1, Q_1^1, Q_1, Q_0)$. Thus, the first order linear expansion of ψ_n is given by:

$$\psi_n - \psi_0 \approx df(Q_0^1, Q_1^1, Q_1, Q_0)(Q_{0n}^1 - Q_0^1, Q_{1n}^1 - Q_1^1, Q_{1n} - Q_1, Q_{0n} - Q_0),$$

which provides the influence curve of ψ_n as an estimator of ψ_0. Thus, we should first determine these derivative of f. We define $\bar{Q}(w) = q_0 Q_1(w) + (1-q_0)Q_0(w)$ and $\bar{Q}(w,1) = q_0 Q_1(w,1) + (1-q_0)Q_0(w,1)$. It follows

$$\begin{aligned}\psi_n - \psi_0 \approx & \sum_w \frac{Q(w,1)q_0^2}{\bar{Q}(w,1)}(Q_{1n} - Q_1)(w) \\ & + \sum_w \frac{Q_1(w,1)q_0(1-q_0)}{\bar{Q}(w,1)}(Q_{0n} - Q_0)(w) \\ & + \sum_w \bar{Q}(w)\frac{q_0}{\bar{Q}(w,1)}(Q_{1n} - Q_1)(w,1) \\ & - \sum_w \bar{Q}(w)\frac{q_0^2 Q_1(w,1)}{\bar{Q}^2(w,1)}(Q_{1n} - Q_1)(w,1) \\ & - \sum_w \bar{Q}(w)\frac{q_0(1-q_0)Q_1(w,1)}{\bar{Q}^2(w,1)}(Q_{0n} - Q_0)(w,1).\end{aligned}$$

Substitution of the empirical distributions for a single observation O_i results in the wished influence curve

$$\begin{aligned}D = & \frac{q_0^2 Q_1(W_1,1)}{\bar{Q}(W_1,1)} + \frac{Q_1(W_2,1)q_0(1-q_0)}{\bar{Q}(W_2,1)} \\ & + \frac{q_0\bar{Q}(W_1)I(A_1=1)}{\bar{Q}(W_1,1)} - \frac{q_0^2\bar{Q}(W_1)I(A_1=1)Q_1(W_1,1)}{\bar{Q}(W_1,1)^2} \\ & - \frac{q_0(1-q_0)\bar{Q}(W_2)I(A_2=1)Q_1(W_2,1)}{\bar{Q}(W_2,1)^2} - c\end{aligned}$$

where c denotes the constant guaranteeing that D has mean zero. Now, we use the following identities:

$$\frac{\bar{Q}(w)}{\bar{Q}(w,1)} = \frac{1}{g_0^*(1 \mid w)}$$

$$\frac{Q_1(w,1)q_0}{\bar{Q}(w,1)} = Q^*(1,w) = P^*(Y=1 \mid A=1, W=w).$$

With these identities it follows

$$\begin{aligned} D &= q_0 \frac{I(A_1=1)}{g_0^*(1 \mid W_1)}(1 - Q^*(1,W_1) + q_0 Q^*(1,W_1) \\ &\quad + (1-q_0)\frac{I(A_2=1)}{g_0^*(1 \mid W_2)}(0 - Q^*(1,W_2) + (1-q_0)Q^*(1,W_2) - c. \end{aligned}$$

Since D needs to have mean zero under P_0 it follows that $c = \psi_0$. □

9.1 Derivation of influence curve of estimator of causal parameter for case-control Design I based on saturated logistic regression model.

Consider the standard logistic regression estimator $\tilde{Q}_n^*$ ignoring the case-control sampling for a saturated logistic regression model and its corresponding estimator $\psi_n(a) = E_{Q_{W,n}^*} Q_{q_0,n}^*(a,W)$ of $\psi_0(a) = E_0 Y_a$. For the sake of presentation, we consider the simple case with $c = 1$, $\bar{d} = 1$. The proof is trivially generalized to general $c, \bar{d}$.

Firstly, we note that $\tilde{Q}_n^*$ solves the score equations

$$0 = \sum_i h(W_{1i}, A_{1i}) - \frac{1}{J}\sum_{j=1}^{J} \frac{\tilde{Q}_n^*}{1-\tilde{Q}_n^*}(W_{2i}^j, A_{2i}^j) h(W_{2i}^j, A_{2i}^j)$$

indexed by arbitrary functions $h(W,A)$.

Now, set $h(W,A) = I_{w,a}(W,A)$ be equal to the indicator that $W = w$ and $A = a$ and we can choose w, a arbitrary. This gives us the equations

$$\frac{\tilde{Q}_n^*(w,a)}{1-\tilde{Q}_n^*(w,a)} = \frac{Q_{1n}(w,a)}{Q_{0n}(w,a)},$$

where $Q_{1n}(w,a) = \frac{1}{n}\sum_i I(W_{1i} = w, A_{1i} = a)$ and $Q_{0n}(w,a) = \frac{1}{nJ}\sum_j \sum_i I(W_{2i}^j = w, A_{2i}^j = a)$ are the empirical distributions of the cases and controls, respectively. It is of interest to note that this empirical relation corresponds with the true known relation

$$\frac{Q_0^*}{1-Q_0^*}(w,a) = \frac{1-q_0}{q_0}\frac{Q_1(w,a)}{Q_0(w,a)}.$$

Thus,

$$\psi_n(a) = \int_w \frac{c_0 Q_{1n}/Q_{0n}(w,a)}{1+c_0 Q_{1n}/Q_{0n}(w,a)} \{q_0 Q_{1n}(w) + (1-q_0)Q_{0n}(w)\} \equiv \Phi(Q_{1n}, Q_{0n}).$$

We also have

$$\psi_0(a) = \int_w \frac{c_0 Q_1/Q_0(w,a)}{1+c_0 Q_1/Q_0(w,a)} \{q_0 Q_1(w) + (1-q_0)Q_0(w)\} \equiv \Phi(Q_1, Q_0).$$

Note that the function $f(x) = c_0 x/(1+c_0 x)$ has derivative $c_0/(1+c_0 x)^2$. So

$$\begin{aligned}\frac{c_0 Q_{1n}/Q_{0n}}{1+c_0 Q_{1n}/Q_{0n}} - \frac{c_0 Q_1/Q_0}{1+c_0 Q_1/Q_0} &\approx c_0/(1+c_0 Q_1/Q_0)^2 (Q_{1n}/Q_{0n} - Q_1/Q_0) \\ &= \frac{c_0}{(1+c_0 Q_1/Q_0)^2}\left\{\frac{1}{Q_0}(Q_{1n}-Q_1) - \frac{Q_1}{Q_0^2}(Q_{0n}-Q_0).\right\}\end{aligned}$$

Thus,

$$\begin{aligned}&\Phi(Q_{1n}, Q_{0n}) - \Phi(Q_1, Q_0) \approx \int_w \frac{c_0}{(1+c_0 Q_1/Q_0)^2}\left\{\frac{1}{Q_0}(Q_{1n}-Q_1) - \frac{Q_1}{Q_0^2}(Q_{0n}-Q_0)\right\}\bar{Q}(w) \\ &+ \int_w \frac{c_0 Q_1/Q_0(w,a)}{1+c_0 Q_1/Q_0(w,a)} \{q_0(Q_{1n}-Q_1)(w) + (1-q_0)(Q_{0n}-Q_0)(w)\},\end{aligned}$$

where we used the notation $\bar{Q}(w) = q_0 Q_1(w) + (1-q_0)Q_0(w)$. Therefore, the influence curve of the estimator $\Phi(Q_{1n}, Q_{0n})$ of $\Phi(Q_1, Q_0)$ is given by

$$\begin{aligned}IC_a(O) &= \frac{c_0}{(1+c_0 Q_1/Q_0)^2}(W_1,a)\frac{\bar{Q}(W_1)}{Q_0(W_1,a)} I(A_1=a) \\ &\quad -\frac{1}{J}\sum_j \frac{c_0}{(1+c_0 Q_1/Q_0)^2}(W_2^j,a)\frac{\bar{Q}(W_2^j)Q_1(W_2^j,a)}{Q_0^2(W_2^j,a)} I(A_2^j=a) \\ &\quad +q_0\frac{c_0 Q_1/Q_0}{1+c_0 Q_1/Q_0}(W_1,a) + (1-q_0)\frac{1}{J}\sum_j \frac{c_0 Q_1/Q_0}{1+c_0 Q_1/Q_0}(W_2^j,a) - \psi_0(a).\end{aligned}$$

Let $R_0^* \equiv Q_1/Q_0$ which is estimated with the logistic regression fit $R_n^* = \tilde{Q}_n^*/(1-\tilde{Q}_n^*)$. Let $K_0(W,A) \equiv \bar{Q}(W)/Q_0(W,A)$. Note,

$$\begin{aligned}IC_a(R_0^*, K_0, \psi_0(a))(O) &= \frac{c_0}{(1+c_0 R_0^*)^2}(W_1,a)K_0(W_1,a)I(A_1=a) \\ &\quad -\frac{1}{J}\sum_j \frac{c_0 R_0^*}{(1+c_0 R_0^*)^2}(W_2^j,a)K_0(W_2^j,a)I(A_2^j=a) \\ &\quad +q_0\frac{c_0 R_0^*}{1+c_0 R_0^*}(W_1,a) + \frac{1}{J}\sum_j (1-q_0)\frac{c_0 R_0^*}{1+c_0 R_0^*}(W_2^j,a) - \psi_0(a).\end{aligned}$$

It follows immediately that

$$E_0 IC(R_0, K, \psi_0(a)) = 0 \text{ for all } K.$$

We will now use another representation of the influence curve IC_a which establishes the wished double robustness. We use

$$\begin{aligned} Q_0^* &= c_0 R_0^*/(1 + c_0 R_0^*) \\ c_0 R_0^* &= Q_0^*/(1 - Q_0^*) \\ K_0(w, a) &= \frac{1 - q_0}{(1 - Q_0^*(w, a)) g_0^*(a \mid w)}. \end{aligned}$$

We have

$$\begin{aligned} IC_a(Q_0^*, g_0^*, \psi_0(a)) &= q_0 \frac{I(A_1 = a)}{g_0^*(a \mid W_1)}(1 - Q_0^*(W_1, a)) \\ &\quad -\frac{1}{J}\sum_j (1 - q_0)\frac{I(A_2^j = a)}{g_0^*(a \mid W_2^j)} Q_0^*(W_2^j, a) \\ &\quad +q_0 Q_0^*(W_1, a) + \frac{1}{J}\sum_j (1 - q_0) Q_0^*(W_2^j, a) - \psi_0(a). \end{aligned}$$

The double robustness for this representation can now be stated as

$$E_0 IC_a(Q^*, g^*, \psi_0(a)) = 0 \text{ if either } g^* = g_0^* \text{ or } Q^* = Q_0^*,$$

and in both cases we need that $g^*(1 \mid W) > 0$ a.e.

Appendix: Derivation of influence curve of particular nonparametric maximum likelihood estimator of causal relative risk for case-control design II.

In this section of the Appendix we consider a causal parameter specified in terms of $r_0(m) = P_0^*(Y = 0, M = m)$, assuming $r_0(m)$ is known. We note that our influence curve results for matched-case-control designs relied on $q_0(1 \mid m)$ being known instead of $r_0(m)$ being known. As a consequence, one now anticipates an additional component (beyond the case-control weighted

influence curve) to the influence curve due to the estimation of $q_0(1 \mid m)$. Indeed, below we derive the influence curve of the nonparametric maximum likelihood estimator and show it involves now an additional term only being a function of M_1.

To start with we derive the identifiability result for EY_1 and subsequently we define the corresponding nonparametric maximum likelihood estimator and derive its influence curve.

We define $Q_1(a,m,w) \equiv P_0^*(M=m, W=w, A=a \mid Y=1)$, and

$$\begin{aligned} Q_0(a,w \mid m) &= P_0^*(A=a, W=w \mid M=m, Y=0) \\ &= \frac{P_0(M_2=m, W_2=w, A_2=a)}{P_0(M=m)} \\ &\equiv \frac{Q_0(a,w,m)}{Q_0(m)}. \end{aligned}$$

We also define $Q_0(w \mid m) = P(W=w \mid M=m, Y=0)$, and, we have $Q_0(w \mid m) = Q_0(w,m)/Q_0(m)$. Let $r_0(m) = P_0(Y=0, M=m)$. We wish to establish an identifiability result of $P(Y_1=1)$ from the distribution P_0 of O: that is, we wish to write $P(Y_1=1)$ as a function of Q_0 and Q_1. Firstly, we note

$$P(Y_1=1) = E_{M,W} P(Y=1 \mid A=1, M, W) = E_W \frac{P(A=1, M, W \mid Y=1) q_0}{P(A=1, M, W)}.$$

Secondly, we note that

$$\begin{aligned} P(A=1, M=m, W=w) &= P(A=1, M=m, W=w \mid Y=1) q_0 \\ &\quad + P(A=1, W=w \mid Y=0, M=m) P(Y=0, M=m). \end{aligned}$$

Finally, we have that

$$\begin{aligned} P(M=m, W=w) &= q_0 p(M=m, W=w \mid Y=1) \\ &\quad + P(Y=0, M=m) p(W=w \mid M=m, Y=0). \end{aligned}$$

Thus, we have shown that

$$\begin{aligned} \psi_0(1) &= P(Y_1=1) \\ &= \sum_{m,w} (q_0 Q_1(m,w) + r_0(m) Q_0(w \mid m)) \frac{Q_1(1,m,w) q_0}{q_0 Q_1(1,m,w) + r_0(m) Q_0(1, w \mid m)} \\ &\equiv \sum_{m,w} \bar{Q}(m,w) \frac{q_0 Q_1(1,m,w)}{\bar{Q}(1,m,w)} \\ &= f(Q_1, Q_0). \end{aligned}$$

We note that this is an identifiability result relying on knowing $r_0(m)$ and q_0 instead of $\bar{q}_0(m)$ and q_0.

Let $Q_{1n}(a,w,m) = \frac{1}{n}\sum_{i=1}^n I(A_{1i}=a, M_{1i}=m, W_{1i}=w)$,

$$\begin{aligned} Q_{0n}(a,w,m) &= \frac{1}{n}\sum_{i=1}^n I(A_{2i}=a, M_{1i}=m, W_{2i}=w) \\ Q_{0n}(m) &= \frac{1}{n}\sum_{i=1}^n I(M_{1i}=m), \end{aligned}$$

and $Q_{0n}(a,w \mid m) = Q_{0n}(a,w,m)/Q_{0n}(m)$. The nonparametric maximum likelihood estimator for the likelihood

$$p_0(O) = Q_1(A_1,M_1,W_1)Q_0(A_2,W_2 \mid M_1),$$

is given by these empirical distribution functions, and the corresponding nonparametric maximum likelihood estimator of EY_1 is thus

$$\psi_n(1) = f(Q_{0n},Q_{1n}).$$

In order to determine the efficient influence curve of EY_1, we will derive the influence curve of the nonparametric maximum likelihood estimator $\psi_n(1)$ as an estimator of $\psi_0(1)$.

We will determine the derivatives of f as a function of $Q_1^1(w,m)$, $Q_0^1(w,m)$, $Q_1(1,w,m)$, $Q_0(1,w,m)$. We note

$$\begin{aligned} &Q_{0n}(a,w \mid m) - Q_0(a,w \mid m) = \tfrac{Q_{0n}(a,w,m)}{Q_{0n}(m)} - \tfrac{Q_0(a,w,m)}{Q_0(m)} \\ &\approx \tfrac{1}{Q_0(m)}(Q_{0n}-Q_0)(a,w,m) - \tfrac{Q_0(a,w,m)}{Q_0(m)^2}(Q_{0n}(m)-Q_0(m) \\ &= \tfrac{1}{n}\sum_{i=1}^n \tfrac{1}{Q_0(m)}\{I(A_{2i}=a,W_{2i}=w,M_{2i}=m) - Q_0(a,w,m)\} \\ &-\tfrac{1}{n}\sum_{i=1}^n \tfrac{Q_0(a,w,m)}{Q_0(m)^2}(I(M_{2i}=m)-Q_0(m)). \end{aligned}$$

Similarly,

$$\begin{aligned} &Q_{0n}(w \mid m) - Q_0(w \mid m) = \tfrac{Q_{0n}(w,m)}{Q_{0n}(m)} - \tfrac{Q_0(w,m)}{Q_0(m)} \\ &\approx \tfrac{1}{Q_0(m)}(Q_{0n}-Q_0)(w,m) - \tfrac{Q_0(w,m)}{Q_0(m)^2}(Q_{0n}(m)-Q_0(m)) \\ &= \tfrac{1}{n}\sum_{i=1}^n \tfrac{1}{Q_0(m)}\{I(W_{2i}=w,M_{2i}=m) - Q_0(w,m)\} - \tfrac{Q_0(w,m)}{Q_0(m)^2}(I(M_{2i}=m)-Q_0(m)). \end{aligned}$$

The first order linear expansion of ψ_n is given by:

$$\psi_n - \psi_0 \approx df(Q_0^1,Q_1^1,Q_1,Q_0)(Q_{0n}^1-Q_0^1, Q_{1n}^1-Q_1^1, Q_{1n}-Q_1, Q_{0n}-Q_0),$$

which provides the influence curve of ψ_n as an estimator of ψ_0. Thus, we should first determine these derivative of f. Recall $\bar{Q}(m,w) = q_0 Q_1(m,w) + r_0(m)Q_0(w \mid m)$ and $\bar{Q}(1,m,w) = q_0 Q_1(1,m,w) + r_0(m)Q_0(1,w \mid m)$. It follows

$$\begin{aligned}
\psi_n - \psi_0 \approx & \sum_{m,w} \frac{Q_1(1,m,w)q_0^2}{\bar{Q}(1,m,w)}(Q_{1n} - Q_1)(m,w) \\
& + \sum_{m,w} \frac{Q_1(1,m,w)q_0 r_0(m)}{\bar{Q}(1,m,w)}(Q_{0n} - Q_0)(w \mid m) \\
& + \sum_{m,w} \bar{Q}(m,w)\frac{q_0}{\bar{Q}(1,m,w)}(Q_{1n} - Q_1)(1,m,w) \\
& - \sum_{m,w} \bar{Q}(m,w)\frac{q_0^2 Q_1(1,m,w)}{\bar{Q}^2(1,m,w)}(Q_{1n} - Q_1)(1,m,w) \\
& - \sum_{m,w} \bar{Q}(m,w)\frac{q_0 r_0(m) Q_1(1,m,w)}{\bar{Q}^2(1,m,w)}(Q_{0n} - Q_0)(1,w \mid m).
\end{aligned}$$

We note $Q_0(m) = P(M_2 = m) = P(M_1 = m) = Q_1(m)$. We also note $r_0(m)/Q_1(m) = \bar{q}_0(m)$. Now, we substitute the empirical distributions and the above empirical approximation for $Q_{0n}(1, w \mid m)$ and $Q_{0n}(w \mid m)$. This yields the influence curve

$$\begin{aligned}
& IC(O_i) = \textstyle\sum_{m,w} q_0 Q^*(1,m,w)\{I(W_{1i} = w, M_{1i} = m) - Q_1(w,m)\} \\
& + \textstyle\sum_{m,w} \bar{q}_0(m) Q^*(1,m,w)\left\{I(W_{2i} = w, M_{2i} = m) - Q_0(w \mid m) I(M_{2i} = m)\right\} \\
& + \textstyle\sum_{m,w} \frac{q_0}{g_0^*(1|m,w)}\left\{I(A_{1i} = 1, M_{1i} = m, W_{1i} = w) - Q_1(1,m,w)\right\} \\
& - \textstyle\sum_{m,w} q_0 \frac{Q^*(1,m,w)}{g_0^*(1|m,w)}\left\{I(A_{1i} = 1, M_{1i} = m, W_{1i} = w) - Q_1(1,m,w)\right\} \\
& - \textstyle\sum_{m,w} \bar{q}_0(m) \frac{Q^*(1,m,w)}{g^*(1|m,w)}\left\{I(A_{2i} = 1, W_{2i} = w, M_{2i} = m)\right\} \\
& - \textstyle\sum_{m,w} \bar{q}_0(m) \frac{Q^*(1,m,w)}{g^*(1|m,w)}\left\{Q_0(1,w \mid m) I(M_{2i} = m)\right\},
\end{aligned}$$

So

$$\begin{aligned}
& IC(O_i) = q_0 Q^*(1,M_1,W_1) - \textstyle\int Q^*(1,m,w) P(M = m, W = w, Y = 1) \\
& + \bar{q}_0(M_1) Q^*(1,M_1,W_2) \\
& - \bar{q}_0(M_1) \textstyle\int_w Q^*(1,M_1,w) P(W = w \mid Y = 0, M = M_1) \\
& + q_0 \frac{I(A_1=1)}{g^*(1|M_1,W_1)} - \textstyle\int Q^*(1,m,w) P(M = m, W = w) \\
& - q_0 I(A_1 = 1) \frac{Q^*(1,M_1,W_1)}{g^*(1|M_1,W_1)} + \textstyle\int Q^{*2}(1,m,w) P(M = m, W = w) \\
& - \bar{q}_0(M_1) \frac{Q^*(1,M_1,W_2)}{g^*(1|M_1,W_2)} I(A_2 = 1) \\
& + \frac{\bar{q}_0(M_1)}{r_0(M_1)} P(M = M_1) \textstyle\int_w Q^*(1,M_1,w)(1 - Q^*(1,M_1,w)) P(W = w \mid M = M_1),
\end{aligned}$$

where we should read $P(W = w \mid M = M_1)$ as $P(W = w \mid M = m)$ evaluated at $m = M_1$. So

$$
\begin{aligned}
&IC(O) = q_0 \left\{ \tfrac{I(A_1=1)}{g^*(1|M_1,W_1)}(1 - Q^*(1, M_1, W_1)) + Q^*(1, M_1, W_1) \right\} \\
&+\bar{q}_0(M_1) \left\{ \tfrac{I(A_2=1)}{g^*(1|M_1,W_2)}(0 - Q^*(1, M_1, W_2)) + Q^*(1, M_1, W_2) \right\} \\
&+ \tfrac{q_0}{P(Y=1|M=M_1)} \left\{ \textstyle\int_w Q^*(1, M_1, w)(Q^*(w, M_1) - Q^*(1, M_1, w)) P(W = w \mid M = M_1) \right\} \\
&- \textstyle\int Q^*(1, m, w) P(M = m, W = w, Y = 1) - \int Q^*(1, m, w) P(M = m, W = w) \\
&+ \textstyle\int Q^{*2}(1, m, w) P(M = m, W = w)
\end{aligned}
$$

References

R.Detrano M. Bobbio E. Gunel A.P. Morise, G.A. Diamond. The effect of disease-prevalence adjustments on the accuracy of a logistic prediction model. *Med Decis Making*, 16:133–142, 1996.

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.

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.

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.

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.

A.M. Molinaro, M.J. van der Laan, D.H. Moore, and K. Kerlikowske. Survival point estimate prediction in matched and non-matched case-control subsample designed studies. Technical report, Division of Biostatistics, University of California, Berkeley, 2005. http://www.bepress.com/ucbbiostat/paper149.

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.

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 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.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 M.L. Petersen. Causal effect models for realistic individualized treatment and intention to treat rules. *International Journal of Biostatistics*, 3(1), 2007.

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.