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

Year 2015 *Paper* 336

Adaptive Pre-specification in Randomized Trials With and Without Pair-Matching

Laura B. Balzer* Mark J. van der Laan†
Maya L. Petersen‡

*Division of Biostatistics, University of California, Berkeley - the SEARCH Consortium, lbbalzer@hsph.harvard.edu

†Division of Biostatistics, University of California, Berkeley - the SEARCH Consortium, laan@berkeley.edu

‡Division of Biostatistics, University of California, Berkeley - the SEARCH Consortium, mayaliv@berkeley.edu

http://biostats.bepress.com/ucbbiostat/paper336

Adaptive Pre-specification in Randomized Trials With and Without Pair-Matching

Laura B. Balzer, Mark J. van der Laan, and Maya L. Petersen

Abstract

In randomized trials, adjustment for measured covariates during the analysis can reduce variance and increase power. To avoid misleading inference, the analysis plan must be pre-specified. However, it is unclear *a priori* which baseline covariates (if any) should be included in the analysis. Consider, for example, the Sustainable East Africa Research in Community Health (SEARCH) trial for HIV prevention and treatment. There are 16 matched pairs of communities and many potential adjustment variables, including region, HIV prevalence, male circumcision coverage and measures of community-level viral load. In this paper, we propose a rigorous procedure to data-adaptively select the adjustment set which maximizes the efficiency of the analysis. Specifically, we use cross-validation to select from a pre-specified library the candidate targeted maximum likelihood estimator (TMLE) that minimizes the estimated variance. For further gains in precision, we also propose a collaborative procedure for estimating the known exposure mechanism. Our small sample simulations demonstrate the promise of the methodology to maximize study power, while maintaining nominal confidence interval coverage. Our procedure is tailored to the scientific question (sample vs. population treatment effect) and study design (pair-matched or not) and alleviates many of the common concerns.

1 Introduction

The objective of a randomized trial is to evaluate the effect of an intervention on the outcome of interest. In this setting, the difference in the average outcomes among the treated units and the average outcomes among the control units provides a simple and unbiased estimate of the intervention effect. Adjusting for measured covariates during the analysis can substantially reduce the estimator's variance and thereby increase study power (e.g. Fisher (1932); Cochran (1957); Cox and McCullagh (1982); Tsiatis et al. (2008); Moore and van der Laan (2009)). Nonetheless, the recommendations on adjustment in randomized trials have been conflicting (Pocock et al., 2002; Hayes and Moulton, 2009; Austin et al., 2010; Kahn et al., 2014; Campbell, 2014). The advice seems to depend on the study design, the unit of randomization, the application, and the sample size. As a result, many researchers are left wondering how to adjust for baseline covariates, if at all.

Consider a trial, where the treatment is randomly allocated to $n/2$ units and the remaining units are assigned to the control. There is a rich literature on locally efficient estimation in this setting (e.g. Tsiatis et al. (2008); Zhang et al. (2008); Rubin and van der Laan (2008); Moore and van der Laan (2009); Shen et al. (2014).) For example, parametric regression can be used to obtain an unbiased and more precise estimate of the intervention effect. Briefly, the outcome is regressed on the exposure and covariates according to a working model. Following Rosenblum and van der Laan (2010), we use "working" to emphasize that the regression function need not be and often is not correctly specified. This working model can include interaction terms and can be linear or non-linear. The estimated coefficients are then used to obtain the predicted outcomes for all units under the treatment and the control. The difference or ratio in the average of the predicted outcomes provides an estimate of the intervention effect.

For continuous outcomes and linear working models without interaction terms, this procedure is known as analysis of covariance (ANCOVA), and the coefficient for the exposure is equal to the estimate of the intervention effect. For binary outcomes, Moore and van der Laan (2009) detailed the potential gains in precision from adjustment via logistic regression for estimating the treatment effect on the absolute or relative scale (i.e. risk difference, risk ratio or odds ratio). Furthermore, the authors showed that parametric maximum likelihood estimation (MLE) was equivalent to targeted maximum likelihood estimation (TMLE) in this setting (van der Laan and Rubin, 2006; van der Laan and Rose, 2011). As a result, the asymptotic properties of the TMLE, including double robustness and asymptotic linearity, hold even if the working model for outcome regression is misspecified. Furthermore, this approach is locally efficient in that the TMLE will achieve the lowest possible variance among a large class of estimators if the working model is correctly specified. Rosenblum and van der Laan (2010) expanded these results for a large class of general linear models. Indeed, the parametric MLE and TMLE can be considered special cases of the double robust estimators of Scharfstein et al. (1999) and semiparametric approaches of Tsiatis et al. (2008); Zhang et al. (2008). For a recent and detailed review of these estimation approaches, we refer the reader to Colantuoni and Rosenblum (2015).

Now consider a pair-matched trial, where the intervention is randomly allocated within the $n/2$ matched pairs. The proposed estimation strategies have been more limited in this setting. Indeed, the perceived "analytical limitations" of pair-matched trials have led some researchers to shy away from this design (Klar and Donner, 1997; Imbens, 2011; Campbell, 2014). As with a completely randomized trial, the unadjusted difference in treatment-specific means provides an unbiased but inefficient estimate of the intervention effect. To include covariates in the analysis and to potentially increase power, Hayes and Moulton (2009) suggested regressing the outcome on the covariates (but not on the exposure) and then contrasting the observed versus predicted

outcomes within matched pairs. Alternatively, TMLE can provide an unbiased and locally efficient approach in pair-matched trials (van der Laan et al., 2012; Balzer et al., 2015b). Specifically, the algorithm can be implemented as if the trial were completely randomized: (1) fit a working model for the mean outcome, given the exposure and covariates, (2) obtain predicted outcomes for all units under the treatment and control, and (3) contrast the average of the predicted outcomes on the relevant scale. Inference, however, must respect the pair-matching scheme (van der Laan et al., 2012; Balzer et al., 2015b).

A common challenge to the both designs is the selection of the covariates for inclusion in the analysis. Many variables are measured prior to implementation of the intervention, and it is difficult to *a priori* specify an appropriate working model. For a completely randomized trial, covariate adjustment will lead to gains in precision if (i) the covariates are predictive of the outcome and (ii) the covariates are imbalanced between treatment groups (e.g. Moore et al. (2011)). Balance is guaranteed as sample size goes to infinity, but rarely seen in practice. Analogously, in a pair-matched trial, covariate adjustment will improve precision if there is an imbalance on predictive covariates after matching.

Limited sample sizes pose an additional challenge to covariate selection. A recent review of randomized clinical trials reported that the median number of participants was 58 with an interquartile range of 27-161 (Califf et al., 2012). Likewise, a recent review of cluster randomized trials reported that the median number of units was 31 with an interquartile range of 13-60 (Selvaraj and Prasad, 2013). In small trials, adjusting for too many covariates can lead to overfitting and inflated Type I error rates (e.g. Moore et al. (2011); Shen et al. (2014); Balzer et al. (2015b)). Finally, *ad hoc* selection of the adjustment set leads to concerns that researchers will go on a "fishing expedition" to find the covariates resulting in most power and again risking inflation of the Type I error rate (e.g. Pocock et al. (2002); Tsiatis et al. (2008); Olken (2015)).

In sum, covariate adjustment in randomized trials can provide meaningful improvements in precision and thereby statistical power. To preserve inference, the working model, including the adjustment variables, must be specified *a priori*. In practice, sample size often limits the size of the adjustment set, and best set is unclear before the trial's conclusion. This results in an important challenge: the need to learn from the data to realize precision gains, but doing so in pre-specified and rigorous way to maintain valid statistical inference.

In this paper, we apply the principle of *empirical efficiency maximization* to data-adaptively select from a pre-specified library the candidate TMLE, which minimizes variance and thereby maximizes the precision of the analysis (Rubin and van der Laan, 2008; van der Laan, 2011). We contribute to the existing methodology by modifying this strategy for pair-matched trials. To our knowledge, such a data-adaptive procedure has not been proposed or implemented for this study design. We further contribute to the literature by collaboratively estimating the exposure mechanism for additional gains in precision (van der Laan and Gruber, 2010; Gruber and van der Laan, 2011). We also generalize the results for estimation and inference of both the population and sample average treatment effects (Balzer et al., 2015a). Our finite sample simulations demonstrate the practical performance with limited numbers of independent units, as is common in early phase clinical trials and in cluster randomized trials. As a motivating example, we discuss the Sustainable East Africa Research in Community Health (SEARCH) study, an ongoing cluster randomized trial for HIV prevention and treatment (NCT01864603) (University of California, San Francisco, 2013). Full `R` code is provided in the Appendix (R Core Team, 2014).

2 Motivating Example and Causal Parameters

SEARCH is a community randomized trial to estimate the effect of immediate and streamlined antiretroviral therapy (ART) on HIV incidence as well as other health, economic and educational outcomes. The trial is being conducted in 32 rural communities in Uganda and Kenya. Extensive baseline characteristics were collected through ethnographic mapping and community-wide censuses. Examples include region, occupational mix, measures of mobility, HIV prevalence and community-level HIV RNA viral load. A subset of these characteristics was used to create the 16 best matched pairs of communities (Balzer et al., 2015b). The intervention was randomized within matched pairs. In treatment communities, HIV testing is expanded, and all individuals testing HIV+ are immediately eligible for ART with enhanced services for initiation, linkage and retention in care. In control communities, all individuals testing HIV+ are eligible for ART, according to in-country guidelines. The primary outcome is the five-year cumulative incidence of HIV and will be measured through longitudinal follow-up. The observed data for a given SEARCH community can be denoted

$$O = (W, A, Y)$$

where W represents the vector of baseline covariates, A represents the intervention assignment, and Y denotes the outcome. Specifically, A is a binary indicator, equalling one if the community was randomized to the treatment and zero if the community was randomized to the control.

In this paper, we consider estimation and inference for the population average treatment effect (PATE) and the sample average treatment effect (SATE). Let $Y(a)$ denote the outcome if possibly contrary-to-fact the unit were assigned intervention-level $A = a$. The causal parameters are a function of the distribution P_X of the full data, comprised of the baseline covariates and the counterfactual outcomes of interest: $X = (W, Y(1), Y(0))$ (Neyman, 1923; Rubin, 1974). Specifically, the PATE is the expected difference in the counterfactual outcomes if all members of the population were assigned the intervention and if all members of that population were assigned the control:

$$\Psi^{\mathcal{P}}(P_X) = E_X\big[Y(1) - Y(0)\big] \tag{1}$$

where the expectation is over the full data distribution P_X. There is one true value of $\Psi^{\mathcal{P}}(P_X)$ for the target population. For the SEARCH trial, the PATE is the expected difference in the counterfactual cumulative incidence of HIV if all communities in the hypothetical target population implemented the test-and-treat strategy and the counterfactual cumulative incidence of HIV if all communities in that target population maintained the standard of care.

The sample parameter is the average difference in the counterfactual outcomes for the study units (Neyman, 1923):

$$\Psi^{\mathcal{S}}(P_X) = \frac{1}{n}\sum_{i=1}^{n}\big[Y_i(1) - Y_i(0)\big] \tag{2}$$

were $Y_i(a)$ denotes the outcome if possibly contrary-to-fact unit i were assigned intervention-level $A = a$. The SATE is data-adaptive; its true value depends on the n units in the sample. The SATE is easily interpretable, responsive to heterogeneity in the intervention effect, and arguably the most relevant when the study units are not representative of a greater population. For the SEARCH trial, the SATE is the average difference in the counterfactual cumulative incidence of HIV under the test-and-treat strategy and under the standard of care for the 32 study communities.

3 Targeted Estimation in a Randomized Trial Without Matching

In this section, we ignore the pair-matching scheme in the SEARCH trial and assume the observed data consist of n independent, identically distribution (i.i.d.) copies of $O = (W, A, Y)$ with some true, but unknown distribution P_0, which factorizes as

$$P_0(O) = P_0(W)P_0(A|W)P_0(Y|A, W).$$

We do not make any assumptions about the common covariate distribution $P_0(W)$ or about the common conditional distribution of the outcome, given the intervention and covariates $P_0(Y|A, W)$. By design, the intervention A is randomized with probability 0.5. Therefore, the exposure mechanism is known: $P_0(A = 1|W) = g_0(1|W) = 0.5$. The statistical model $\mathcal{M}$, describing the set of possible observed data distributions, is semiparametric.

Since the intervention is randomized, we can easily identify the population effect $\Psi^{\mathcal{P}}(P_X)$ (Eq. 1) from the observed data distribution. Our statistical estimand is the difference in the expected outcome, given the treatment and covariates, and the expected outcome, given the control and covariates, averaged (standardized) with respect to the covariate distribution in the population (Robins, 1986):

$$\begin{aligned}\Psi(P_0) &= E_0\big[E_0(Y|A = 1, W) - E_0(Y|A = 0, W)\big] \\ &= E_0\big[\bar{Q}_0(1, W) - \bar{Q}_0(0, W)\big]\end{aligned}$$

where $\bar{Q}_0(A, W) = E_0(Y|A, W)$ denotes the conditional mean outcome, given the exposure and covariates. As discussed in the introduction, there are many algorithms available for unbiased and locally efficient estimation of this statistical parameter in a randomized trial (e.g. Tsiatis et al. (2008); Zhang et al. (2008); Rubin and van der Laan (2008); Moore and van der Laan (2009); Shen et al. (2014)). Throughout, our focus is on TMLE, a general methodology for the construction of double robust, semiparametric efficient substitution estimators (van der Laan and Rubin, 2006; van der Laan and Rose, 2011).

A TMLE for population effect $\Psi^{\mathcal{P}}(P_X)$ (Eq. 1) also serves as a consistent and asymptotically linear estimator of the sample effect $\Psi^{\mathcal{S}}(P_X)$ (Eq. 2) (Balzer et al., 2015a). The estimator can be implemented in three steps.

Step 1. Initial estimation: Estimate the expected outcome, given the exposure and covariates $\bar{Q}_0(A, W) = E_0(Y|A, W)$. We could rely on a pre-specified parametric working model (as discussed above) or implement a more data-adaptive approach (as discussed below).

Step 2. Targeting: Update the initial estimator $\bar{Q}_n(A, W)$.

i. Calculate the "clever" covariate based on the known or estimated exposure mechanism $g_n(A|W) = P_n(A|W)$:

$$H_n(A, W) = \left(\frac{\mathbb{I}(A = 1)}{g_n(1|W)} - \frac{\mathbb{I}(A = 0)}{g_n(0|W)}\right)$$

ii. If the outcome is continuous and unbounded, run linear regression of the outcome Y on the covariate $H_n(A, W)$ with the initial estimator as offset. Plug in the estimated coefficient ϵ_n to yield the targeted update: $\bar{Q}^*_n(A, W) = \bar{Q}_n(A, W) + \epsilon_n H_n(A, W)$.

iii. If the outcome is binary or bounded in $[0,1]$*, run logistic regression of the outcome Y on the covariate $H_n(A,W)$ with the $\text{logit}(x) = log\{x/(1-x)\}$ of the initial estimator $\bar{Q}_n(A,W)$ as offset. Plug in the estimated coefficient ϵ_n to yield the targeted update: $\bar{Q}_n^*(A,W) = \text{expit}\big\{\text{logit}[\bar{Q}_n(A,W)] + \epsilon_n H_n(A,W)\big\}$, where expit is the inverse-logit.

Step 3. Parameter estimation: Obtain the predicted outcomes for all observations under the treatment $\bar{Q}_n^*(1,W)$ and control $\bar{Q}_n^*(0,W)$. Average the difference in predicted outcomes:

$$\Psi_n(\bar{Q}_n^*) = \frac{1}{n}\sum_{i=1}^{n}\left[\bar{Q}_n^*(1,W_i) - \bar{Q}_n^*(0,W_i)\right]$$

If the initial estimator for $\bar{Q}_0(A,W)$ is based on a working regression model with an intercept and a main term for the exposure and if the exposure mechanism is treated as known (i.e. not estimated), then the updating step can be skipped (Rosenblum and van der Laan, 2010). Further precision, however, can be attained by using a data-adaptive algorithm for initial estimation of $\bar{Q}_0(A,W)$ and by estimating the exposure mechanism $g_0(A|W)$ (van der Laan and Robins, 2003).

Under regularity conditions, the standardized estimator converges to a normal distribution with mean 0 and variance given by the variance of its influence curve, divided by sample size n. The influence curve for the TMLE of the population parameter (PATE) can be estimated from the observed data distribution by

$$D_n^{\mathcal{P}}(g_n,\bar{Q}_n^*)(O) = \left(\frac{\mathbb{I}(A=1)}{g_n(1|W)} - \frac{\mathbb{I}(A=0)}{g_n(0|W)}\right)\big(Y - \bar{Q}_n^*(A,W)\big) + \bar{Q}_n^*(1,W) - \bar{Q}_n^*(0,W) - \Psi_n(\bar{Q}_n^*) \tag{3}$$

(van der Laan and Rose, 2011). The influence curve for the TMLE of the sample parameter (SATE) can be *conservatively* estimated from the observed data distribution by

$$D_n^{\mathcal{S}}(g_n,\bar{Q}_n^*)(O) = \left(\frac{\mathbb{I}(A=1)}{g_n(1|W)} - \frac{\mathbb{I}(A=0)}{g_n(0|W)}\right)\big(Y - \bar{Q}_n^*(A,W)\big) \tag{4}$$

(Balzer et al., 2015a). For the SATE, there is no variance contribution from the covariate distribution, which is considered fixed. Asymptotically, the SATE will often be estimated with more precision than the PATE (Neyman, 1923; Imbens, 2004). However, as shown by Balzer et al. (2015a) and in the simulations that follow, the gains in precision from specifying the SATE as the target of inference can be attenuated in small trials, because this influence curve-based variance estimator is conservative.

3.1 Adaptive Pre-specified Approach for Step 1. Initial Estimation

Consider again the SEARCH trial for HIV prevention and treatment. Recall that the outcome Y is the five-year cumulative incidence of HIV and bounded between 0 and 1. Suppose that we want to use logistic regression for initial estimation of the expected outcome, given the exposure and measured covariates $\bar{Q}_0(A,W)$. It is unclear *a priori* which covariates should be included in the working model and in what form. For example, baseline HIV prevalence is a known predictor of the outcome and may be imbalanced between the treatment and control groups. Likewise,

*In greater generality, the logistic fluctuation can also be used for a continuous outcome that is bounded in $[a,b]$ by first applying the following transformation to the outcome: $Y^* = (Y-a)/(b-a)$. For further details, see Gruber and van der Laan (2010).

there might be substantial heterogeneity in the treatment effect by region and allowing for an interaction between region and the intervention may reduce the variance of the TMLE. Including all the covariates and the relevant interactions in the working model is likely to result in overfitting and misleading inference. To facilitate selection between candidate initial estimators and thereby candidate TMLEs, we propose the following cross-validation selector.

First, we propose a library of candidate working models for initial estimation of the conditional mean outcome $\bar{Q}_0(A,W)$. This library should be pre-specified in the protocol or the analysis plan. A possible library could consist of the following logistic regression working models:

$$\begin{aligned}
&\text{logit}[\bar{Q}^{(a)}(A,W,\beta)] = \beta_0 + \beta_1 A \\
&\text{logit}[\bar{Q}^{(b)}(A,W,\beta)] = \beta_0 + \beta_1 A + \beta_2 W1 \\
&\text{logit}[\bar{Q}^{(c)}(A,W,\beta)] = \beta_0 + \beta_1 A + \beta_2 W2 + \beta_3 A \times W2
\end{aligned}$$

where, for example, $W1$ denotes baseline prevalence and $W2$ denotes region. Of course, there are many more candidate algorithms, and we are considering this simple set for pedagogic purposes. We also note that the first working model corresponds to the unadjusted estimator.

Second, we need to pre-specify a loss function to measure the performance of candidate estimators. Following the principle of empirical efficiency maximization (van der Laan, 2011), we propose using the variance of the estimated influence curve of the TMLE for the parameter of interest. Specifically, if the target of inference is the population effect, our loss function is

$$\mathcal{L}^{\mathcal{P}}(\bar{Q}) = Var\big[D_n^{\mathcal{P}}(g_0,\bar{Q})\big],$$

and if the target of inference is the sample effect, our loss function is

$$\mathcal{L}^{\mathcal{S}}(\bar{Q}) = Var\big[D_n^{\mathcal{S}}(g_0,\bar{Q})\big].$$

For SATE, this corresponds to the L2 squared error loss function $\mathcal{L}^{\mathcal{S}}(\bar{Q}) = \big(Y - \bar{Q}(A,W)\big)^2$.

Next, we need to pre-specify our cross-validation scheme, used to generate an estimate of the expected loss (i.e. the "risk") for each of the candidate estimators. For generality, we present V-fold cross-validation, where the data are randomly split into V partitions, called "folds", of size $\approx n/V$. To respect the limited sample sizes common in early phase clinical trials and in cluster randomized trials, leave-one-out cross-validation may be appropriate. Leave-one-out cross-validation corresponds with $V = n$-fold cross-validation, where each fold corresponds to one observation. The cross-validation procedure for initial estimation of the conditional mean $\bar{Q}_0(A,W)$ can be implemented as follows.

i. For each fold $v = \{1,\ldots,V\}$ in turn,

   a. Set the observation(s) in fold v to be the validation set and the remaining observations to be the training set.
   b. Fit each algorithm for estimating $\bar{Q}_0(A,W)$ using only data in the training set. For the above library, we would run logistic regression of the outcome Y on the exposure A and covariates W, according to the working model. Denote the initial regression fits as $\bar{Q}_n^{(a)}(A,W)$, $\bar{Q}_n^{(b)}(A,W)$ and $\bar{Q}_n^{(c)}(A,W)$, respectively.
   c. For each algorithm, use the estimated fit to predict the outcome(s) for the observation(s) in the validation set under the treatment and the control. For the first algorithm, for example, we would have $\bar{Q}_n^{(a)}(1,W_k)$ and $\bar{Q}_n^{(a)}(0,W_k)$ for observation O_k in the validation set.

ii. For each algorithm, estimate the risk with the sample variance of the cross-validated estimate of the influence curve. If the target of inference is the PATE, our risk estimates would be

$$\text{CV-risk}^{\mathcal{P},(a)} = Var_n\big[D_n^{\mathcal{P}}\big(g_0, \bar{Q}_n^{(a)}\big)\big]$$
$$\text{CV-risk}^{\mathcal{P},(b)} = Var_n\big[D_n^{\mathcal{P}}\big(g_0, \bar{Q}_n^{(b)}\big)\big]$$
$$\text{CV-risk}^{\mathcal{P},(c)} = Var_n\big[D_n^{\mathcal{P}}\big(g_0, \bar{Q}_n^{(c)}\big)\big]$$

where Var_n denotes the sample variance and we are treating the exposure mechanism as known: $g_0(A|W) = 0.5$. If instead the target of inference is SATE, our risk estimates would be

$$\text{CV-risk}^{\mathcal{S},(a)} = Var_n\big[D_n^{\mathcal{S}}(g_0, \bar{Q}_n^{(a)})\big]$$
$$\text{CV-risk}^{\mathcal{S},(b)} = Var_n\big[D_n^{\mathcal{S}}\big(g_0, \bar{Q}_n^{(b)}\big)\big]$$
$$\text{CV-risk}^{\mathcal{S},(c)} = Var_n\big[D_n^{\mathcal{S}}\big(g_0, \bar{Q}_n^{(c)}\big)\big]$$

iii. Select the algorithm with the smallest cross-validated risk.

The selected working model is then used for initial estimation of $\bar{Q}_0(A, W)$ in Step 1 of the TMLE algorithm. Specifically, we would re-fit the selected algorithm $\bar{Q}_n(A, W)$ using all the data. Since our library was limited to parametric working models and the exposure mechanism was treated as known, the updating step (Step 2) can be skipped. In other words, the chosen estimator is already targeted $\bar{Q}_n(A, W) = \bar{Q}_n^*(A, W)$ and can be used for Step 3 parameter estimation.

4 Targeted Estimation in a Randomized Trial With Matching

Recall the pair-matching scheme briefly described in Section 2 for the SEARCH trial. First, the potential study units were selected. Then baseline covariates, such as region, occupational mix and measures of migration, were collected. A matching algorithm was applied to the baseline covariates of candidate units to create the best 16 matched pairs. The intervention was randomized within the resulting pairs, and the outcome will be measured with longitudinal follow-up. This pair-matching scheme is considered to be *adaptive*, because the resulting matched pairs are a function of the baseline covariates of all the candidate units (van der Laan et al., 2012; Balzer et al., 2015b). This design has also been called "nonbipartite matching" and "optimal multivariate matching" (Greevy et al., 2004; Zhang and Small, 2009; Lu et al., 2011).

The adaptive design creates a dependence in the data. Since the construction of the matched pairs is a function of the baseline covariates of all n study units, the observed data do not consist of $n/2$ i.i.d. paired observations, as current practice sometimes assumes (e.g. Klar and Donner (1997); Freedman et al. (1997); Campbell et al. (2007); Hayes and Moulton (2009)). Instead, we have n dependent copies of $O = (W, A, Y)$. Nonetheless, there is a lot of conditional independence in the data. Mainly, once we consider the baseline covariates of the study units as fixed, we recover $n/2$ conditionally independent units:

$$\bar{O}_j = \big(O_{j1}, O_{j2}\big) = \big((W_{j1}, A_{j1}, Y_{j1}), (W_{j2}, A_{j2}, Y_{j2})\big)$$

where the index $j = 1, \ldots, n/2$ denotes the partitioning of the candidate units $\{1, \ldots n\}$ into matched pairs according to similarity in their baseline covariates $(W_1, \ldots, W_n)$. Throughout subscripts $j1$ and $j2$ index the observations within matched pair j. The conditional distribution of the

observed data, given the baseline covariates of the study units, factorizes as

$$P_0(O_1,\ldots,O_n|W_1,\ldots W_n) = \prod_{j=1}^{n/2} P_0(A_{j1},A_{j2}|W_1,\ldots,W_n)P_0(Y_{j1}|A_{j1},W_{j1})P_0(Y_{j2}|A_{j2},W_{j2})$$
$$= 0.5\prod_{j=1}^{n/2} P_0(Y_{j1}|A_{j1},W_{j1})P_0(Y_{j2}|A_{j2},W_{j2})$$

where the second line follows from randomization of the intervention within matched pairs. For estimation and inference of the PATE, we need to assume that each community's baseline covariates W_i are independently drawn from some common distribution $P_0(W)$. For estimation and inference of the SATE, this assumption on the covariate distribution can be weakened. (See the Appendix A for further details.)

Despite the dependence in the data, a TMLE for the population effect (PATE) can be implemented as if the sample were n i.i.d. units (van der Laan et al., 2012). In Step 1, we obtain an initial estimator of $\bar{Q}_0(A,W)$ with an *a priori*-specified parametric working model or with a more data-adaptive method. In Step 2, we target the initial estimator $\bar{Q}_n(A,W)$ by using information in the known or estimated exposure mechanism. Finally in Step 3, we obtain the predicted outcomes for all observations under the treatment $\bar{Q}_n^*(1,W)$ and the control $\bar{Q}_n^*(0,W)$, and then take the sample average of the difference in these predicted outcomes. Furthermore, the variance of the TMLE can be estimated by treating the sample as n i.i.d. units. In other words, inference can be based on the sample variance of the estimated influence curve in the non-matched trial $D_n^{\mathcal{P}}$ (Eq. 3), divided by n (van der Laan et al., 2012). This variance estimator ignores any gains in precision from pair-matching and will be conservative under reasonable assumptions. A less conservative variance estimator can be obtained by accounting for the potential correlations of the residuals within matched pairs:

$$\rho_n(\bar{Q}_n^*)(\bar{O}_j) = \frac{1}{n/2}\sum_{j=1}^{n/2}\big(Y_{j1}-\bar{Q}_n^*(A_{j1},W_{j1})\big)\big(Y_{j2}-\bar{Q}_n^*(A_{j2},W_{j2})\big) \tag{5}$$

(van der Laan et al., 2012). An estimate of the asymptotic variance of the TMLE is then given by the sample variance of $D_n^{\mathcal{P}}$ minus $2\rho_n$, all divided by n.

In a pair-matched trial, a TMLE for the population effect is also a consistent and asymptotically linear estimator of the sample effect. The proof is given in Appendix A. Furthermore, the influence curve for the TMLE of the SATE can be *conservatively* estimated by

$$\bar{D}_n^{\mathcal{S}}(g_n,\bar{Q}_n^*)(\bar{O}_j) = \frac{1}{2}\left[D_n^{\mathcal{S}}(g_n,\bar{Q}_n^*)(O_{j1}) + D_n^{\mathcal{S}}(g_n,\bar{Q}_n^*)(O_{j2})\right] \tag{6}$$

where $D_n^{\mathcal{S}}(g_n,\bar{Q}_n^*)(O)$ is the estimated influence curve for observation O in the non-matched trial (Eq. 4). The proof is given in Appendix B. Inference can be based on the sample variance of the estimated (paired) influence curve $\bar{D}_n^{\mathcal{S}}$, divided by $n/2$. If we order observations within matched pairs such that first corresponds to the intervention ($A_{j1}=1$) and the second to the control ($A_{j2}=0$) and treat the exposure mechanism as known $g_0(A|W)=0.5$, we have

$$\bar{D}_n^{\mathcal{S}}(g_0,\bar{Q}_n^*)(\bar{O}_j) = \big(Y_{j1}-\bar{Q}_n^*(1,W_{j1})\big) - \big(Y_{j2}-\bar{Q}_n^*(0,W_{j2})\big)$$

In this setting, the sample variance of the pairwise differences in residuals, divided by $n/2$, provides a conservative variance estimator.

4.1 Adaptive Pre-specified Approach for Step 1. Initial Estimation

By balancing intervention groups with respect to baseline determinants of the outcome, pair-matching increases the efficiency of the study (e.g. Imai et al. (2009); van der Laan et al. (2012); Balzer et al. (2015b)). Nonetheless, residual imbalance on the baseline predictors often remains, and adjusting for these covariates during the analysis can further increase efficiency. In the SEARCH trial, for example, the matched pairs were created before baseline HIV prevalence was measured. As a result, there is likely to be variation across pairs in baseline prevalence, which is a known driver of HIV incidence. Adjusting for baseline prevalence during the analysis is likely to reduce the variance of the TMLE and result in a less conservative variance estimator. Unfortunately, it is unclear *a priori* whether adjusting for prevalence will yield more power than than adjusting for other baseline covariates, such as male circumcision coverage or measures of community-level HIV RNA viral load. With only 16 (conditionally) independent units, we are limited as to the size of the adjustment set. Adjusting for too many covariates can result in over-fitting. As before, we want to data-adaptively select the candidate TMLE (i.e. working regression model), which maximizes the empirical efficiency.

The data-adaptive procedure for Step 1 initial estimation of conditional mean outcome $\bar{Q}_0(A, W)$, outlined in Sec. 3.1 for a non-matched trial, can be modified for a pair-matched trial. As before, we need to pre-specify our library of candidate estimators, our measure of performance and the cross-validation scheme. We can use the same library of candidate working models for initial estimation of the conditional mean outcome $\bar{Q}_0(A, W)$. For the loss function, however, we want to use the estimated variance of the TMLE under pair-matching. To elaborate, consider the loss function for the SATE in a non-matched trial. Minimizing the sum of squared residuals (i.e. minimizing the variance of $D_n^{\mathcal{S}}$ (Eq. 4)) targets the conditional mean outcome $\bar{Q}_0(A, W)$. As a result, the algorithm could select a working model adjusting for a covariate that is highly predictive of the outcome but on which we matched perfectly. In the SEARCH trial, for example, communities were paired within region, because HIV incidence is expected to be highly heterogeneous across regions. Therefore, minimizing the empirical variance of $D_n^{\mathcal{S}}$ might lead to the selection of the candidate TMLE with main terms for the intervention and region. This selection would not improve the precision of the analysis over the unadjusted algorithm. Instead, we want to select the TMLE minimizing the conservative estimator of the variance in the pair-matched design. Thereby, our loss functions for the PATE and SATE are

$$\mathcal{L}^{\mathcal{P}}(\bar{Q}) = Var\left[D_n^{\mathcal{P}}(g_0, \bar{Q}) - 2\rho_n(\bar{Q})\right]$$
$$\mathcal{L}^{\mathcal{S}}(\bar{Q}) = Var\left[\bar{D}_n^{\mathcal{S}}(g_0, \bar{Q})\right]$$

respectively. Finally, the pair should be treated as the unit of (conditional) independence in the cross-validation scheme. In other words, when the data are split into V-folds, the pairing should be preserved. In small trials, leave-one-pair-out cross-validation may be appropriate. With these modifications, we can implement the cross-validation scheme, outlined in Sec. 3.1, to data-adaptively select the candidate working model, which minimizes the estimated variance of the TMLE in a pair-matched trial.

5 Collaborative Estimation of the Exposure Mechanism

Even though the intervention A is randomized with balanced allocation, estimating the known exposure mechanism $g_0(A|W)$ can increase the precision of the analysis (van der Laan and Robins, 2003). As before, we want to respect the study design (i.e. pair-matched or not) as well as adjust

for a covariate only if its inclusion improves the empirical efficiency. For example, we may not want to include a covariate that is imbalanced between the intervention groups (i.e. predictive of A) but not predictive of the outcome. Likewise, if a given covariate (e.g. $W1$) was included in the working model for $\bar{Q}_0(A, W)$, further adjusting for this covariate when estimating the exposure mechanism may not increase precision. To this end, we propose to incorporate the Collaborative TMLE (C-TMLE) approach into our algorithm (van der Laan and Gruber, 2010; Gruber and van der Laan, 2011).

5.1 Adaptive Pre-specified Approach for Step 2. Targeting

First, we propose a library of candidate estimators of the exposure mechanism $g_0(A|W)$. As before, this library should be pre-specified in the protocol or analysis plan. A possible library could consist of the following logistic regression working models:

$$\begin{aligned}
\text{logit}[g^{(a)}(W, \beta)] &= \beta_0 \\
\text{logit}[g^{(b)}(W, \beta)] &= \beta_0 + \beta_1 W1 \\
\text{logit}[g^{(c)}(W, \beta)] &= \beta_0 + \beta_1 W2
\end{aligned}$$

where, for example, $W1$ is baseline prevalence and $W2$ is region. Each algorithm would yield a different update to a given initial estimator of the conditional mean outcome $\bar{Q}_n(A, W)$, selected by the data-adaptive procedure for Step 1 (Sec. 3.1 and 4.1). In other words, each candidate estimator of $g_0(A|W)$ results in a different targeted estimator $\bar{Q}^*_n(A, W)$. We also note that the first working model corresponds to the unadjusted estimator.

To choose between candidate algorithms, we need to pre-specify a loss function. As before, we propose using the estimated variance of the TMLE, appropriate for the scientific question (i.e. population or sample effect) and study design (i.e. pair-matched or not). Finally, we need to pre-specify our cross-validation scheme, used to obtain an honest measure of risk and to reduce the potential for over-fitting. As before, we present V-fold cross-validation, where the data are partitioned into V folds of size $\approx n/V$. If matching was used, the partitioning should preserve the pairs. The cross-validation selector for collaborative estimation of the exposure mechanism can be implemented as follows.

i. For each fold $v = \{1, \ldots, V\}$ in turn,

   a. Set the observation(s) in fold v to be the validation set and the remaining observations to be the training set.

   b. Using only data in the training set, fit each algorithm for estimating the exposure mechanism. For the above library, we would run logistic regression of the exposure A on the covariates W, according to the working model. Denote the estimated exposure mechanisms as $g^{(a)}_n(A|W)$, $g^{(b)}_n(A|W)$ and $g^{(c)}_n(A|W)$, respectively.

   c. For each algorithm, use the estimated fit of the exposure mechanism to target the initial estimator $\bar{Q}_n(A, W)$, also fit with the training set.

   d. For each algorithm, obtain targeted predictions of the outcome(s) for the observation(s) in the validation set under the treatment and the control. For the first algorithm, for example, we would have $\bar{Q}^{(a),*}_n(1, W_k)$ and $\bar{Q}^{(a),*}_n(0, W_k)$ for observation O_k in the validation set.

ii. For each algorithm, estimate the risk with the cross-validated variance estimator, appropriate for the target parameter and study design.

iii. Select the algorithm with the smallest cross-validated risk.

The chosen estimator is then used for targeting in Step 2 of the TMLE algorithm.

6 Obtaining Inference

In summary, we have proposed the following data-adaptive TMLE to maximize the precision and power of a randomized trial.

> Step 1. Initial estimation of the conditional mean outcome with the working model $\bar{Q}_n(A, W)$, which was data-adaptively selected to maximize the empirical efficiency of the analysis (Sec. 3.1 for a non-matched trial and Sec. 4.1 for a matched trial).
>
> Step 2. Targeting the initial estimator using the estimated exposure mechanism $g_n(A|W)$, which was data-adaptively selected to further maximize the empirical efficiency of the analysis (Sec. 5.1).
>
> Step 3. Obtaining a point estimate by averaging the difference in the targeted predicted outcomes:
>
> $$\Psi_n(\bar{Q}_n^*) = \frac{1}{n}\sum_{i=1}^{n}\left[\bar{Q}_n^*(1, W_i) - \bar{Q}_n^*(0, W_i)\right]$$

We now need a variance estimator that accounts for the selection process. For this, we propose using a cross-validated variance estimator. As before, the data are split into validation and training sets, respecting the unit of (conditional) independence. The selected TMLE is fit using the data in the training set and used to estimate the influence curve[†] for the observation(s) in the validation set. The step-by-step instructions are given in the Appendix C. The sample variance of the cross-validated estimate of the influence curve can then be used for hypothesis testing and the construction of Wald-type confidence intervals. For trials with a limited number of independent units, the Student's t-distribution is an appropriate alternative to the standard normal distribution.

7 Small Sample Simulations

We present the following simulation studies to demonstrate (1) implementation of the proposed methodology, (2) the potential gains in precision and power from data-adaptive estimation of the conditional mean outcome, (3) the additional gains in precision and power from collaborative estimation of the exposure mechanism, and (4) maintenance of nominal confidence interval coverage. All simulations were conducted in R v3.1.2 (R Core Team, 2014).

7.1 Study 1

For each unit $i = \{1, \ldots, n\}$, we generated the nine baseline covariates by drawing from a multivariate normal with mean 0 and variance 1. The correlation between the first three covariates $\{W1, W2, W3\}$ and between the second three covariates $\{W4, W5, W6\}$ was 0.5, while the correlation between the remaining covariates $\{W7, W8, W9\}$ was 0. The exposure A was randomized such that the treatment allocation was balanced overall. For the non-matched trial, we randomly

[†]For the TMLE of the population effect in a pair-matched trial, we also need a cross-validated estimate of the correction term ρ_n (Eq. 5.) This term is a function of the residuals, which can be estimated for each pair in the validation set based on targeted estimator $\bar{Q}_n^*(A, W)$, fit with the training set.

assigned the intervention to $n/2$ units and the control to the remaining $n/2$ units. For the pair-matched trial, we used the non-bipartite matching algorithm `nbpMatch` to pair units on covariates $\{W1, \ldots, W6\}$ (Lu et al., 2012). The exposure A was randomized within the resulting matched pairs. Recall A is a binary indicator, equalling 1 if the unit was assigned the treatment and 0 if the unit was assigned the control. For each unit, the outcome Y was then generated as

$$Y = 0.4A + 0.25(W1 + W2 + W4 + W5 + U_Y) + 0.25A(W1 + U_Y)$$

where U_Y was drawn from a standard normal. We also generated the counterfactual outcomes $Y(1)$ and $Y(0)$ by intervening to set $A = a$. To reflect the limited sample sizes common in early phase clinical trials and in cluster randomized trials, we selected a sample size of $n = 40$. This resulted in $n/2 = 20$ conditionally independent units in the pair-matched trial.

For each study design (non-matched or matched), this data generating process was repeated 2,500 times. Recall that the sample effect $\Psi^{\mathcal{S}}(P_X)$ (Eq. 2) is data-adaptive parameter; its value changes with each new selection of units. Thereby, for each repetition, the SATE was calculated as the sample average of the difference in the counterfactual outcomes. The SATE ranged from 0.23 to 0.60 with a mean of 0.40. In contrast, the population effect $\Psi^{\mathcal{P}}(P_X)$ (Eq. 1) is constant and was calculated by averaging the difference in the counterfactual outcomes over a population of 900,000 units. The true value of the PATE was 0.40.

We compared the performance of the unadjusted estimator to TMLE with various approaches to covariate adjustment. Specifically, we implemented the TMLE algorithm, where the initial estimation of the conditional mean outcome $\bar{Q}_0(A, W)$ was based on a linear working model with main terms for the intervention A and irrelevant covariate $W9$ and where the exposure mechanism was treated as known: $g_0(A|W) = 0.5$. This approach was equivalent to standard maximum likelihood estimation (MLE) and represented the unfortunate scenario where the researcher pre-specified adjustment for a covariate that was not predictive of the outcome.

We also implemented a TMLE with the data-adaptive approach for Step 1 initial estimation of the conditional mean outcome (Sec. 3.1 and 4.1). Our library consisted of 10 working linear regression models, each with an intercept, a main term for the exposure A and a main term for one baseline covariate: $\{\emptyset, W1, \ldots, W9\}$, where $\emptyset$ corresponds to the unadjusted estimator. Our loss function was the estimated variance of the TMLE, appropriate for the target parameter and study design. We chose the candidate working model with the lowest estimated risk, based on leave-one-out cross-validation for the non-matched trial and leave-one-pair-out cross-validation for the matched trial. We also implemented Collaborative-TMLE (C-TMLE), which couples the data-adaptive approach for Step 1 initial estimation of the conditional mean outcome (Sec. 3.1 and 4.1) with the data-adaptive approach for Step 2 targeting (Sec. 5.1). For the latter, our library of candidates to estimate the exposure mechanism consisted of 10 working logistic regression models, each with an intercept and a main term for one baseline covariate: $\{\emptyset, W1, \ldots, W9\}$. The same loss function and cross-validation scheme were used for C-TMLE.

For the unadjusted estimator and the MLE, inference was based on the estimated influence curve. For the data-adaptive TMLEs, inference was based on the cross-validated estimate of the influence curve (Sec. 6). We assumed the standardized estimator followed the Student's t-distribution with $n - 2 = 38$ degrees of freedom for the non-matched trial and with $n/2 - 1 = 19$ degrees of freedom for the matched trial.

7.1.1 Results

Table 1 illustrates the performance of the estimators over the 2,500 simulated data sets. Specifically, we show the mean squared error (MSE), the relative MSE (rMSE), the average standard error

estimate $\hat{\sigma}$, the attained power and the 95% confidence interval coverage. As expected, matching improved efficiency. The MSE of the unadjusted estimator, for example, was approximately 2 times larger in the non-matched trial than in the pair-matched trial. Furthermore, for the pair-matched trial, targeting the sample effect, as opposed to the population effect, resulted in substantial gains in attained power: 38% with the unadjusted estimator for the PATE and 53% with the same estimator for the SATE. For the non-matched trial, targeting the sample parameter increased efficiency, but did not directly translate into increased power due to the conservative variance estimator for the SATE.

	PATE					**SATE**				
	MSE[a]	rMSE[b]	$\hat{\sigma}$[c]	Pow[d]	Cov[e]	MSE[a]	rMSE[b]	$\hat{\sigma}$[c]	Pow[d]	Cov[e]
Non-Matched										
Unadj.	6.3E-2	1.00	0.25	0.36	0.95	6.0E-2	1.05	0.25	0.36	0.95
MLE	6.6E-2	0.96	0.25	0.37	0.94	6.3E-2	1.00	0.25	0.37	0.95
TMLE	4.4E-2	1.44	0.20	0.51	0.94	4.2E-2	1.52	0.20	0.49	0.95
C-TMLE	4.2E-2	1.51	0.20	0.52	0.95	3.9E-2	1.63	0.20	0.50	0.96
Matched										
Unadj.	3.3E-2	1.93	0.22	0.38	0.99	3.0E-2	2.11	0.18	0.53	0.97
MLE	3.4E-2	1.84	0.22	0.38	0.98	3.2E-2	1.99	0.18	0.54	0.96
TMLE	2.5E-2	2.53	0.18	0.56	0.98	2.4E-2	2.63	0.16	0.68	0.95
C-TMLE	2.4E-2	2.66	0.18	0.58	0.98	2.3E-2	2.78	0.15	0.70	0.95

[a] Mean squared error: the bias (average deviation between the point estimate and sample-specific true value) - squared plus the variance
[b] Relative MSE: the MSE of the unadjusted estimator for the PATE in a non-matched trial relative to (divided by) the MSE of another estimator
[c] Average standard error estimate, based on the estimated influence curve
[d] Attained power: proportion of times the false null hypothesis was rejected
[e] Confidence interval (CI) coverage: proportion of times the true value was contained in the 95% CI

Table 1: Summary of estimator performance for Simulation 1. The rows denote the study design and the estimator: unadjusted, MLE adjusting for $W9$, TMLE with data-adaptive selection of the initial estimator, and Collaborative-TMLE (C-TMLE) with data-adaptive selection of the initial estimator paired with data-adaptive estimation of the exposure mechanism.

In all scenarios, the TMLE with data-adaptive selection of the initial estimator of $\bar{Q}_0(A, W)$ improved precision over the unadjusted estimator and the MLE. Collaborative estimation of the exposure mechanism $g_0(A|W)$ led to further gains in precision. Consider, for example, estimation of the PATE in a trial without matching. The MSE of the unadjusted estimator was 1.44 times larger than the TMLE and 1.51 times larger than the C-TMLE. The attained power was 36%, 51% and 52%, respectively. Furthermore, the precision of the MLE, adjusting for the irrelevant covariate $W9$, was worse than the other estimators in all scenarios. This demonstrates the potential peril of relying on one pre-specified adjustment variable. As a second example, consider the attained power to detect that the SATE was different from zero in the pair-matched trial. We would have 53% power with the unadjusted estimator and 54% power with the MLE, adjusting for the irrelevant covariate $W9$. By incorporating the cross-validation selector for initial estimation of $\bar{Q}_0(A, W)$, the TMLE achieved 68% power. By further incorporating collaborative estimation of the exposure mechanism $g_0(A|W)$, the C-TMLE achieved 70% power. Overall, the greatest efficiency was achieved with C-TMLE for the SATE in the pair-matched trial. Indeed, the MSE of the unadjusted estimator for the population parameter in the trial without matching was nearly 3 times larger than the MSE of the C-TMLE for the sample effect in the pair-matched trial. Throughout the confidence interval

		Working model for initial estimation of $\bar{Q}_0(A,W)$									
Adjustment variable		$\emptyset$	$W1$	$W2$	$W3$	$W4$	$W5$	$W6$	$W7$	$W8$	$W9$
PATE	non-matched	0	57	19	1	11	11	1	0	0	0
	matched	0	54	20	1	12	12	1	0	0	0
SATE	non-matched	0	57	19	1	11	11	1	0	0	0
	matched	0	38	21	5	15	15	3	0	1	1
		Working model for estimation of $g_0(A\|W)$									
Adjustment variable		$\emptyset$	$W1$	$W2$	$W3$	$W4$	$W5$	$W6$	$W7$	$W8$	$W9$
PATE	non-matched	79	2	3	2	3	3	2	2	2	2
	matched	29	9	10	9	9	9	8	5	6	6
SATE	non-matched	77	2	3	2	3	4	2	2	3	2
	matched	16	10	11	10	11	10	10	7	7	7

Table 2: For Simulation 1, the proportion of times a covariate was selected in the working linear regression model for initial estimation of $\bar{Q}_0(A,W)$ and in the working logistic regression model for collaborative estimation of the exposure mechanism $g_0(A|W)$.

coverage was maintained near or above the nominal rate of 95%.

Further insight into the efficiency gains with the proposed TMLE and C-TMLE is provided by Table 2, which shows the proportion of times a working model was selected for initial estimation of the conditional mean outcome and for collaborative estimation of the exposure mechanism. When targeting the PATE, the selection for $\bar{Q}_0(A,W)$ was similar with and without pair-matching. This was not surprising, because our measure of performance (i.e. the loss function) was the estimated variance of the TMLE, and the variance estimator in a pair-matched trial is given by the estimated variance in the non-matched trial minus a correction term ρ_n, which was close to 0. When targeting the SATE, however, the selection procedure was more optimized to the study design. For example, the working model with main terms for the intervention and $W1$ was selected in 57% of the studies without matching and in only 38% of the studies with matching. Instead, working models adjusting for other predictive covariates were selected more frequently. Furthermore, the collaborative procedure for estimation of the exposure mechanism was able to identify settings where the no adjustment would yield the greatest gains in efficiency. Specifically, the unadjusted estimator $g_n(A|W) = 0.5$ was selected in nearly 80% of the studies without matching and in less than 30% of the studies with matching.

7.2 Study 2

For the second simulation study, we increased the complexity of the data-generating process and reduced the sample size to $n = 30$. As before, we generated nine baseline covariates from a multivariate normal with mean 0, variance 1 and the same correlation structure. We also generated a binary variable R, equalling 1 with probability 0.5 and equalling -1 with probability 0.5. The final covariate Z was generated as a function of these baseline covariates and random noise U_Z:

$$Z = R \times \text{expit}(W1 + W4 + W7 + 0.5U_Z)$$

where the expit is the inverse of the logit function and U_Z was drawn independently from a standard normal. As before, the intervention A was randomized with balanced allocation. For a non-matched trial, the treatment was randomly assigned to $n/2$ units and the control to the remaining $n/2$ units. For the pair-matched trial, we used the non-bipartite matching algorithm `nbpMatch` to explore two

		correlation 0.5			correlation 0.5			correlation 0			
	R	$W1$	$W2$	$W3$	$W4$	$W5$	$W6$	$W7$	$W8$	$W9$	Z
Parents of covariate Z	✓	✓			✓			✓			
Parents of the outcome Y	✓		✓			✓			✓		✓
Matching set 1	✓										
Matching set 2	✓		✓			✓			✓		

Table 3: For Simulation 2, the relationships between baseline covariates and the outcome as well as the adaptive pair-matching schemes.

matching sets (Lu et al., 2012). In the first, units were matched on R, a baseline covariate strongly impacting Z. In the second, units were matched on $\{R, W2, W5, W8\}$. The intervention A was randomized within the matched pairs. For each unit, the outcome Y was then generated as

$$Y = \text{expit}[0.75A + 0.5(W2 + W5 + W8) + 1.5Z + 0.25U_Y + 0.75A(W2 - W5) + 0.5AZ]/7.5$$

where U_Y was drawn from a standard normal. Thereby, the outcome was a continuous variable bounded in $[0, 1]$ (e.g. a proportion). We also generated the counterfactual outcomes $Y(1)$ and $Y(0)$ by intervening to set $A = a$. For each study design, this data generating process was repeated 2,500 times. The SATE and PATE were calculated as before. The SATE ranged from 0.25% to 3.1% with a mean of 1.6%. The true value of the PATE was 1.6%. Table 3 depicts the relationship between the baseline covariates and the outcome as well as the adaptive pair-matching schemes.

We compared the same algorithms: the unadjusted estimator, the MLE adjusting for the irrelevant covariate $W9$, the TMLE with data-adaptive initial estimation of the conditional mean outcome, and the C-TMLE pairing data-adaptive initial estimation of the conditional mean outcome with data-adaptive targeting. Our library for initial estimation of the conditional mean outcome $\bar{Q}_0(A, W)$ consisted of 12 working logistic regression models, each with an intercept and a main term for the exposure A and a main term for one candidate adjustment variable $\{\emptyset, R, W1, \ldots, W9, Z\}$. Our library for collaborative estimation of the exposure mechanism $g_0(A|W)$ included 12 working logistic regression models, each with an intercept and a main term for one candidate adjustment variable: $\{\emptyset, R, W1, \ldots, W9, Z\}$. We used the same measure of performance and cross-validation scheme. As before, inference was based on the estimated influence curve for the unadjusted estimator and the MLE and on the cross-validated estimate of the influence curve for the data-adaptive TMLEs (Sec. 6). We assumed the standardized estimator followed the Student's t-distribution with $n - 2 = 28$ degrees of freedom for the non-matched trial and with $n/2 - 1 = 14$ degrees of freedom for the matched trial.

7.2.1 Results

The results for the second simulation study are given in Table 4 and largely echoed the above findings. Pair-matching, even on a single covariate (i.e. matching set 1), improved the precision of the analysis. Targeting the sample effect instead of the population effect further improved efficiency. Allowing for data-adaptive selection of the working model for initial estimation of $\bar{Q}_0(A, W)$ yielded even greater precision, and the most efficient analysis was with C-TMLE. Indeed, the MSE of the unadjusted estimator for the PATE in the non-matched trial was nearly 5 times higher than the MSE of the C-TMLE when matching on predictive covariates (i.e. matching set 2). This resulted in over 30% more power to detect the intervention effect.

	PATE					SATE				
	MSE[a]	rMSE[b]	$\hat{\sigma}^c$	Pow[d]	Cov[e]	MSE[a]	rMSE[b]	$\hat{\sigma}^c$	Pow[d]	Cov[e]
Non-matched										
Unadj.	1.7E-4	1.00	0.013	0.21	0.94	1.5E-4	1.16	0.013	0.21	0.96
MLE	1.8E-4	0.97	0.013	0.23	0.94	1.5E-4	1.12	0.013	0.23	0.95
TMLE	1.1E-4	1.54	0.010	0.33	0.93	9.0E-5	1.92	0.010	0.31	0.96
C-TMLE	1.1E-4	1.57	0.010	0.34	0.93	8.6E-5	1.99	0.010	0.32	0.97
Match Set 1										
Unadj.	1.2E-4	1.43	0.012	0.21	0.96	9.9E-5	1.74	0.011	0.28	0.97
MLE	1.3E-4	1.34	0.011	0.23	0.96	1.1E-4	1.63	0.011	0.29	0.96
TMLE	9.9E-5	1.74	0.009	0.33	0.95	7.5E-5	2.30	0.009	0.38	0.96
C-TMLE	9.6E-5	1.79	0.009	0.36	0.94	7.5E-5	2.31	0.008	0.43	0.95
Match Set 2										
Unadj.	6.4E-5	2.70	0.011	0.18	0.99	4.5E-5	3.85	0.009	0.36	0.99
MLE	7.1E-5	2.43	0.011	0.21	0.99	5.2E-5	3.28	0.009	0.37	0.99
TMLE	5.1E-5	3.39	0.009	0.31	0.99	3.5E-5	4.86	0.008	0.46	0.98
C-TMLE	5.1E-5	3.35	0.009	0.35	0.98	3.6E-5	4.78	0.007	0.52	0.98

[a]Mean squared error: the bias (average deviation between the point estimate and sample-specific true value) - squared plus the variance
[b]Relative MSE: the MSE of the unadjusted estimator for the PATE in a non-matched trial relative to (divided by) the MSE of another estimator
[c]Average standard error estimate, based on the estimated influence curve
[d]Attained power: proportion of times the false null hypothesis was rejected
[e]Confidence interval (CI) coverage: proportion of times the true value was contained in the 95% CI

Table 4: Summary of estimator performance for Simulation 2. The rows denote the study design and the estimator: unadjusted, MLE adjusting for $W9$, TMLE with data-adaptive selection of the initial estimator, and Collaborative-TMLE (C-TMLE) with data-adaptive selection of the initial estimator paired with data-adaptive estimation of the exposure mechanism.

For these simulations, there was a notable impact of parameter specification on estimator performance. We first focus on the estimation of the PATE and then on estimation of the SATE. When the population effect was the target of inference, the gains in attained power from pair-matching were attenuated despite the gains in MSE. This was likely due to the slight underestimation of the standard error in the non-matched trial and overestimation in the pair-matched trial. Indeed, the confidence interval coverage in the non-matched trial was less than nominal (93-94%), while the coverage when matching well (i.e. set 2) approached 100%. For this set of simulations, the correction factor ρ_n (Eq. 5) used in variance estimation for the pair-matched design was approximately 0. As a result, the variance estimator in the pair-matched trial was quite conservative, and the cross-validation selection scheme was more optimized for the non-matched trial. The latter point is evidenced by Table 5, which shows the proportion of times a candidate working model was selected. The logistic regression model adjusting for R was selected for initial estimation of $\bar{Q}_0(A, W)$ in 10% of the studies without matching and in 8% of the studies when matching well on R (i.e. set 1). Furthermore, when matching on several covariates (i.e. set 2), the selection of working models for $\bar{Q}_0(A, W)$ was very similar to the selection in the non-matched trial.

In contrast, when estimating the SATE, smaller MSE translated to greater attained power, while maintaining nominal, if not conservative, confidence interval coverage. For example, the attained power of the TMLE was 31% in the non-matched trial, 38% when matching on a single covariate and 46% when matching on several covariates. Likewise, the attained power of the C-TMLE was

		Selected in the working model for initial estimation of $\bar{Q}_0(A,W)$											
Adjustment variable		$\emptyset$	R	$W1$	$W2$	$W3$	$W4$	$W5$	$W6$	$W7$	$W8$	$W9$	Z
PATE	non-matched	0	10	0	16	1	0	0	0	0	3	0	70
	matched set1	0	8	1	30	1	0	0	0	0	4	0	56
	matched set2	0	8	0	16	1	0	0	0	0	2	0	72
SATE	non-matched	0	11	1	18	1	0	0	0	0	3	0	67
	matched set1	0	2	3	55	3	1	1	1	1	11	1	21
	matched set2	0	6	2	30	3	2	1	2	1	7	2	44
		Selected in the working model for estimation of $g_0(A\|W)$											
Adjustment variable		$\emptyset$	R	$W1$	$W2$	$W3$	$W4$	$W5$	$W6$	$W7$	$W8$	$W9$	Z
PATE	non-matched	79	2	2	2	2	1	1	1	2	3	2	2
	matched set1	37	9	5	6	5	4	3	3	5	5	5	13
	matched set2	25	10	4	9	6	5	9	5	4	8	4	11
SATE	non-matched	78	2	2	3	2	1	1	1	2	3	2	2
	matched set1	25	9	5	7	6	6	5	5	6	7	6	11
	matched set2	13	10	6	12	6	5	9	6	6	10	5	11

Table 5: For Simulation 2, the proportion of times a covariate was selected in the working logistic regression model for initial estimation of $\bar{Q}_0(A,W)$ and in the working logistic regression model for collaborative estimation of the exposure mechanism $g_0(A|W)$.

32% in the non-matched trial, 43% in the trial pair-matching on a single covariate and 52% in trial matching on several covariates. From Table 5, we see that the working model adjusting for R was selected for initial estimation of $\bar{Q}_0(A,W)$ in 11% of the studies without matching and only in 2% of the studies when matching well on R (i.e. set 1). In the latter, more weight was given to other predictive baseline covariates, such as $W2$ and $W8$.

8 Discussion

This paper builds on the rich history of covariate adjustment in randomized trials (e.g. Fisher (1932); Cochran (1957); Cox and McCullagh (1982); Tsiatis et al. (2008); Zhang et al. (2008); Moore et al. (2011); Yuan et al. (2012); Shen et al. (2014); Colantuoni and Rosenblum (2015)). In particular, Rubin and van der Laan (2008) proposed the principle of *empirical efficiency maximization* as a strategy to select the estimator of $\bar{Q}_0(A,W)$ that minimized the empirical variance of the estimated efficient influence curve. Their procedure, however, relied on solving a weighted nonlinear least squares problem. Our approach only requires researchers to take the sample variance of the estimated influence curve. More recently, van der Laan and Gruber (2010) proposed collaborative estimation of the exposure mechanism to achieve the greatest bias reduction in the targeting step of TMLE in a observational study. In randomized trials, there is no risk of bias from regression model misspecification (e.g. Rosenblum and van der Laan (2010)). Thereby, the collaborative approach, implemented here, serves only to increase precision by estimating the known exposure mechanism. To our knowledge, this is the first research into C-TMLE in a randomized trial setting. Most recently, van der Laan (2011) suggested selection of the candidate (C-)TMLE based on minimizing the estimated variance of its influence curve. Our paper generalizes this scheme for estimation and inference of both the population and sample average treatment effects in randomized trials with and without pair-matching.

Our simulations illustrate the performance of the proposed procedure in realistically-sized (i.e.

small) trials. In particular, with only 15 (conditionally) independent units, our procedure was able to identify the optimal working model for initial estimation of $\bar{Q}_0(A, W)$ from a library of 12 candidates as well as for collaborative estimation of $g_0(A|W)$ from a library of 12 candidates, while maintaining close to nominal confidence interval coverage. The simulations also indicated the most efficient approach was estimating the sample effect with C-TMLE in pair-matched trial. Indeed, this approach was nearly 5 more efficient than targeting the population effect with the unadjusted estimator in the non-matched trial. Thereby, our procedure dispels the common concerns of "analytical limitations" to pair-matched trials (e.g. Klar and Donner (1997); Imbens (2011); Campbell (2014)).

There are several areas of future work. First, our library of candidate estimators was limited to simple parametric working models. This choice was made for pedagogic purpose and to avoid over-fitting in small trials. In larger trials, we can expand the library to include working models with multiple adjustment variables and interactions as well as selection procedures (e.g. stepwise regression) and semiparametric algorithms. Future work will involve simulations to evaluate the methodology in larger trials. The application to matched triplets, as opposed to matched pairs, should be straightforward. However, the impact of adaptive stratification on estimation and inference merits additional consideration. Finally, we focused on two causal parameters: the population and sample average treatment effects. TMLE is a general methodology for the construction of double robust, semiparametric, efficient substitution estimators for a wide range of parameters. Our proposed strategy for covariate selection should extend to other causal parameters, such as the conditional average treatment effect, the average treatment effect among the treated, and the natural direct effect.

Overall, we proposed a general strategy to increase power in randomized trials. Specifically, we used cross-validation to select the candidate TMLE that optimized the efficiency of the analysis. Since the step-by-step algorithm (including the library definition) was pre-specified, there was no risk of bias or misleading inference from *ad hoc* analytic decisions. In other words, we have proposed a black box procedure to data-adaptively select the most powerful analysis. Furthermore, including the unadjusted estimator as a candidate obviates the need for guidelines on whether or not to adjust (e.g. Moore et al. (2011); Colantuoni and Rosenblum (2015)). Finally, our procedure is tailored to the scientific question (population vs. sample effect) and study design (with or without pair-matching). Decisions about whether to adjust and how to adjust are made with a rigorous and principled approach, removing some of the "human art" from statistics.

9 Acknowledgements

The authors would like to acknowledge and thank the entire SEARCH collaboration for their helpful comments and discussion. This work was supported, in part, by the National Institute Of Allergy And Infectious Diseases of the National Institutes of Health under award numbers R01-AI074345 and UM1AI069502. The content is solely the responsibility of the authors and does not necessarily represent the officials views of the National Institutes of Health. Maya Petersen is a recipient of a Doris Duke Clinical Scientist Development Award.

References

A. Abadie and G. Imbens. Simple and bias-corrected matching estimators for average treatment effects. Technical Report 283, NBER technical working paper, 2002.

P.C. Austin, A. Manca, M. Zwarensteina, D.N. Juurlinka, and M.B. Stanbrook. A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. *Journal of Clinical Epidemiology*, 63 (142-153), 2010.

L.B. Balzer, M.L. Petersen, and M.J. van der Laan. Targeted estimation and inference of the sample average treatment effect. Technical Report 334, Division of Biostatistics, University of California at Berkeley, March 2015a. URL `http://biostats.bepress.com/ucbbiostat/paper334/`.

L.B. Balzer, M.L. Petersen, M.J. van der Laan, and the SEARCH Consortium. Adaptive pair-matching in randomized trials with unbiased and efficient effect estimation. *Statistics in Medicine*, 34(6):999–1011, 2015b.

R.M. Califf, D.A. Zarin, J.M. Kramer, R.E. Sherman, L.H. Aberle, and A. Tasneem. Characteristics of clinical trials registered in ClinicalTrials.gov, 2007-2010. *JAMA*, 307(17):1838–1847, 2012.

M.J. Campbell. Cluster randomized trials. In W. Ahrens and I. Pigeot, editors, *Handbook of Epidemiology, 2nd edition*. Springer, 2014.

M.J. Campbell, A. Donner, and N. Klar. Developments in cluster randomized trials and *Statistics in Medicine*. *Statistics in Medicine*, 26(1):2–19, 2007. doi: 10.1002/sim.2731.

W.G. Cochran. Analysis of covariance: its nature and uses. *Biometrics*, 13:261–281, 1957.

E. Colantuoni and M. Rosenblum. Leveraging prognostic baseline variables to gain precision in randomized trials. Technical Report 263, Johns Hopkins University, Dept. of Biostatistics Working Papers, February 2015. URL `http://biostats.bepress.com/jhubiostat/paper263`.

D.R. Cox and P. McCullagh. Some aspects of analysis of covariance. *Biometrics*, 38(3):541–561, 1982.

R.A. Fisher. *Statistical methods for research workers*. Oliver and Boyd Ltd., Edinburgh, 4th edition, 1932.

L.S. Freedman, M.H. Gail, S.B. Green, D.K. Corle, and The COMMIT Research Group. The Efficiency of the Matched-Pairs Design of the Community Intervention Trial for Smoking Cessation (COMMIT). *Controlled Clinical Trials*, 18(2):131–139, 1997. doi: 10.1016/S0197-2456(96)00115-8.

R. Greevy, B. Lu, J.H. Silber, and P. Rosenbaum. Optimal multivariate matching before randomization. *Biostatistics*, 5(2):263–275, 2004. doi: 10.1093/biostatistics/5.2.263.

S. Gruber and M.J. van der Laan. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *The International Journal of Biostatistics*, 6(1):Article 26, 2010. doi: 10.2202/1557-4679.1260.

S. Gruber and M.J. van der Laan. C-TMLE of an Additive Point Treatment Effect. In M.J. van der Laan and S. Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York Dordrecht Heidelberg London, 2011.

R.J. Hayes and L.H. Moulton. *Cluster Randomised Trials*. Chapman & Hall/CRC, Boca Raton, 2009.

K. Imai, G. King, and C. Nall. The essential role of pair matching in cluster-randomized experiments, with application to the Mexican Universal Health Insurance Evaluation. *Statistical Science*, 24(1):29–53, 2009. doi: 10.1214/08-STS274.

G.W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: a review. *Review of Economics and Statistics*, 86(1):4–29, 2004. doi: 10.1162/003465304323023651.

G.W. Imbens. Experimental design for unit and cluster randomized trials. Technical report, NBER Technical Working Paper, 2011.

B.C. Kahn, V. Jairath, C.J. Doré, and T.P. Morris. The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials*, 15(139):1–7, 2014.

N. Klar and A. Donner. The merits of matching in community intervention trials: a cautionary tale. *Statistics in Medicine*, 16(15):1753–1764, 1997. doi: 10.1002/(SICI)1097-0258(19970815)16:15<1753::AID-SIM597>3.0.CO;2-E.

B. Lu, R. Greevy, X. Xu, and C. Beck. Optimal Nonbipartite Matching and its Statistical Applications. *American Statistician*, 65(1):21–30, 2011. doi: 10.1198/tast.2011.08294.

B. Lu, R. Greevy, and C. Beck. *nbpMatching: functions for non-bipartite optimal matching*, 2012. URL `http://CRAN.R-project.org/package=nbpMatching`. R package version 1.3.6.

K.L. Moore and M.J. van der Laan. Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *Statistics in Medicine*, 28(1):39–64, 2009. doi: 10.1002/sim.3445.

K.L. Moore, R. Neugebauer, T. Valappil, and M.J. van der Laan. Robust extraction of covariate information to improve estimation efficiency in randomized trials. *Statistics in Medicine*, 30(19):2389–2408, 2011.

J. Neyman. Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes (In Polish). English translation by D.M. Dabrowska and T.P. Speed (1990). *Statistical Science*, 5:465–480, 1923.

B.A. Olken. Pre-analysis plans in economics. Technical report, Massachusetts Institute of Technology Department of Economics, 2015. URL `http://economics.mit.edu/files/10399`.

S.J. Pocock, S.E. Assmann, L.E. Enos, and L.E. Kasten. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine*, 21(19):2917–2930, 2002. doi: 10.1002/sim.1296.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org`.

J.M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods–application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:1393–1512, 1986.

M. Rosenblum and M.J. van der Laan. Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. *The International Journal of Biostatistics*, 6(1):Article 13, 2010. doi: 10.2202/1557-4679.1138.

Daniel B. Rubin and M.J. van der Laan. Empirical efficiency maximization: improved locally efficient covariate adjustment in randomized experiments and survival analysis. *The International Journal of Biostatistics*, 4(1):Article 5, 2008. doi: 10.2202/1557-4679.1084.

D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350.

D.O. Scharfstein, A. Rotnitzky, and J.M. Robins. Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models (with Rejoiner). *Journal of the American Statistical Association*, 94(448):1096–1120 (1135–1146), 1999. doi: 10.2307/2669930.

S. Selvaraj and V. Prasad. Characteristics of cluster randomized trials: Are they living up to the randomized trial? *JAMA Internal Medicine*, 173(23):313, 2013.

C. Shen, X. Li, and L. Li. Inverse probability weighting for covariate adjustment in randomized studies. *Statistics in Medicine*, 33:555–568, 2014.

A.A. Tsiatis, M. Davidian, M. Zhang, and X. Lu. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine*, 27(23):4658–4677, 2008. doi: 10.1002/sim.3113.

University of California, San Francisco. Sustainable East Africa Research in Community Health (SEARCH). ClinicalTrials.gov, 2013. URL `http://clinicaltrials.gov/show/NCT01864603`.

M. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York Dordrecht Heidelberg London, 2011.

M.J. van der Laan. Appendix A.19: Efficiency maximization and TMLE. In M.J. van der Laan and S. Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York Dordrecht Heidelberg London, 2011.

M.J. van der Laan and S. Gruber. Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics*, 6(1), 2010.

M.J. van der Laan and J.M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer-Verlag, New York Berlin Heidelberg, 2003.

M.J. van der Laan and D.B. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):Article 11, 2006. doi: 10.2202/1557-4679.1043.

M.J. van der Laan, L.B. Balzer, and M.L. Petersen. Adaptive Matching in Randomized Trials and Observational Studies. *Journal of Statistical Research*, 46(2):113–156, 2012.

S. Yuan, H.H. Zhang, and M. Davidian. Variable selection for covariate-adjusted semiparametric inference in randomized clinical trials. *Statistics in Medicine*, 31:3789–3804, 2012.

K. Zhang and D.S. Small. Comment: The essential role of pair matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation. *Statistical Science*, 25(1):59–64, 2009. doi: 10.1214/09-STS274B.

M. Zhang, A.A. Tsiatis, and M. Davidian. Improving Efficiency of Inferences in Randomized Clinical Trials Using Auxiliary Covariates. *Biometrics*, 64(3):707–715, 2008. doi: 10.1111/j.1541-0420.2007.00976.x.

Appendix A: The TMLE is an asymptotically linear estimator of the SATE in an Adaptive Pair-Matched Trial

In this Appendix we first review the asymptotic linearity results of Balzer et al. (2015b) for estimation and inference of the the statistical parameter corresponding to the conditional average treatment effect (CATE) (Abadie and Imbens, 2002):

$$\Psi_0^{\mathcal{C}}(P_0) = \frac{1}{n}\sum_{i=1}^{n}\left[\bar{Q}_0(1,W_i) - \bar{Q}_0(0,W_i)\right].$$

We then provide a theorem showing that the TMLE for the SATE is asymptotically normal in a trial with adaptive pair-matching, which results in $n/2$ conditionally independent copies of $\bar{O}_j = \big(O_{j1}, O_{j2}\big) = \big((W_{j1}, A_{j1}, Y_{j1}), (W_{j2}, A_{j2}, Y_{j2})\big)$.

As discussed in Balzer et al. (2015b), the TMLE for conditional estimand $\Psi_0^{\mathcal{C}}(P_0)$ is defined by the following substitution estimator:

$$\Psi_n(P_n) = \frac{1}{n}\sum_{i=1}^{n}\left[\bar{Q}_n^*(1,W_i) - \bar{Q}_n^*(0,W_i)\right]$$

where $\bar{Q}_n^*(A,W)$ denotes the targeted estimator. Under the following assumptions, the TMLE for $\Psi_0^{\mathcal{C}}(P_0)$ is asymptotically linear:

$$\Psi_n(P_n) - \Psi_0^{\mathcal{C}}(P_0) = \frac{1}{n/2}\sum_{j=1}^{n/2}\bar{D}^{\mathcal{C}}(\bar{Q},\bar{Q}_0,g_0)(\bar{O}_j) + o_P(\sqrt{n/2})$$

with influence curve

$$\begin{aligned}
\bar{D}^{\mathcal{C}}(\bar{Q},\bar{Q}_0,g_0)(\bar{O}_j) &= \frac{1}{2}\Big[D^{\mathcal{C}}(\bar{Q},\bar{Q}_0,g_0)(O_{j1}) + D^{\mathcal{C}}(\bar{Q},\bar{Q}_0,g_0)(O_{j2})\Big] \\
D^{\mathcal{C}}(\bar{Q},\bar{Q}_0,g_0)(O_i) &= \left(\frac{\mathbb{I}(A_i=1)}{g_0(A_i)} - \frac{\mathbb{I}(A_i=0)}{g_0(A_i)}\right)\big(Y_i - \bar{Q}(A_i,W_i)\big) \\
&\quad - \Big[\big(\bar{Q}_0(1,W_i) - \bar{Q}(1,W_i)\big) - \big(\bar{Q}_0(0,W_i) - \bar{Q}(0,W_i)\big)\Big]
\end{aligned}$$

where $\bar{Q}(A,W)$ denotes the limit of the targeted estimator of the conditional mean function $\bar{Q}_0(A,W)$ and where the marginal probability of being assigned the treatment or the control is known: $g_0(A) = P_0(A) = 0.5$ (Balzer et al., 2015b). Specifically, we assume

- Uniform bound: Assume $\sup_{\bar{Q}\in\mathcal{F}}\sup_O \mid \left(\frac{\mathbb{I}(A_i=1)}{g_0(A_i)} - \frac{\mathbb{I}(A_i=0)}{g_0(A_i)}\right)\big(Y_i - \bar{Q}(A_i,W_i)\mid < M < \infty$ where $\mathcal{F}$ is the set of multivariate real valued functions so that $\bar{Q}_n^*$ is an element of $\mathcal{F}$ with probability 1 and where the second supremum is over a set that contains the support of each O_i.
- Convergence of variances: Assume that for a specified $\{\sigma^{2,\mathcal{C}}(\bar{Q}) : \bar{Q}\in\mathcal{F}\}$, for any $\bar{Q}\in\mathcal{F}$, $\frac{1}{n/2}\sum_{j=1}^{n/2} P_0^n\bar{D}^{\mathcal{C}}(\bar{Q},\bar{Q}_0,g_0)^2 \to \sigma^{2,\mathcal{C}}(\bar{Q})$ a.s (i.e., for almost every $(W^n, n\geq 1)$). Throughout $P_0^n f = E_0[f|W^n]$ denotes the conditional expectation of a function f of $O^n = (O_1,\ldots,O_n)$, given the vector of baseline covariates $W^n = (W_1,\ldots,W_n)$. We will relax this assumption below.

- Convergence of $\bar{Q}_n^*$ to some limit: For any $\bar{Q}_1, \bar{Q}_2 \in \mathcal{F}$, we define $\sigma_n^2(\bar{Q}_1 - \bar{Q}_2) = \frac{1}{n/2}\sum_{j=1}^{n/2} P_0^n\{\bar{D}^{\mathcal{C}}(\bar{Q}_1, \bar{Q}_0, g_0) - \bar{D}^{\mathcal{C}}(\bar{Q}_2, \bar{Q}_0, g_0)\}^2$. Assume that for a particular $\bar{Q} \in \mathcal{F}$, $\sigma_n^2(\bar{Q}_n^* - \bar{Q}) \to 0$ in probability as $n \to \infty$.
- Entropy condition: Let $\mathcal{F}^d = \{f_1 - f_2 : f_1, f_2 \in \mathcal{F}\}$. Let $N(\epsilon, \sigma_n, \mathcal{F}^d)$ be the covering number of the class $\mathcal{F}^d$ w.r.t norm/dissimilarity $\parallel f \parallel = \sigma_n(f)$. Assume that the class $\mathcal{F}$ satisfies $\lim_{\delta_n \to 0} \int_0^{\delta_n} \sqrt{\log N(\epsilon, \sigma_n, \mathcal{F}^d)} d\epsilon = 0$.

Theorem 1. *Let W denote the measured baseline covariates, A the intervention assignment and Y the outcome. A randomized trial with adaptive pair-matching results in $n/2$ conditionally independent copies of paired random variable*

$$\bar{O}_j = \big(O_{j1}, O_{j2}\big) = \big((W_{j1}, A_{j1}, Y_{j1}), (W_{j2}, A_{j2}, Y_{j2})\big)$$

where index $j = \{1, \ldots, n/2\}$ denotes the partitioning of the study units $\{1, \ldots n\}$ into matched pairs according to similarity on their baseline covariates $W^n = (W_1, \ldots, W_n)$. Our target of inference is the sample average treatment effect (SATE) (Neyman, 1923):

$$\Psi^{\mathcal{S}}(P_X) = \frac{1}{n}\sum_{i=1}^{n} Y_i(1) - Y_i(0)$$

where P_X denotes the distribution of the full data $X = \big(W, Y(1), Y(0)\big)$. Under the above conditions, the TMLE $\Psi_n(P_n) = \frac{1}{n}\sum_{i=1}^{n} \bar{Q}_n^*(1, W_i) - \bar{Q}_n^*(0, W_i)$ is an asymptotically linear estimator of the SATE:

$$\Psi_n(P_n) - \Psi^{\mathcal{S}}(P_X) = \frac{1}{n/2}\sum_{j=1}^{n/2} \bar{D}^{\mathcal{S}}(\bar{Q}, \bar{Q}_0, g_0)(\bar{X}_j, \bar{O}_j) + o_P(\sqrt{n/2})$$

with influence curve

$$\bar{D}^{\mathcal{S}}(\bar{Q}, \bar{Q}_0, g_0)(\bar{X}_j, \bar{O}_j) = \bar{D}^{\mathcal{C}}(\bar{Q}, \bar{Q}_0, g_0)(\bar{O}_j) - \bar{D}^{\mathcal{F}}(\bar{Q}_0)(\bar{X}_j, \bar{O}_j)$$

where $\bar{Q}(A, W)$ denotes the limit of the targeted estimator of the conditional mean function $\bar{Q}_0(A, W)$ and where the marginal probability of being assigned the treatment or the control is known $g_0(A) = P_0(A)$.

The first component $\bar{D}^{\mathcal{C}}(\bar{O}_j)$ is the influence curve for the TMLE targeting the conditional estimand $\Psi_0^{\mathcal{C}}(P_0) = \frac{1}{n}\sum_{i=1}^{n}\big[\bar{Q}_0(1, W_i) - \bar{Q}_0(0, W_i)\big]$ in a trial with adaptive pair-matching:

$$\begin{aligned}
\bar{D}^{\mathcal{C}}(\bar{Q}, \bar{Q}_0, g_0)(\bar{O}_j) &= \frac{1}{2}\Big[D^{\mathcal{C}}(\bar{Q}, \bar{Q}_0, g_0)(O_{j1}) + D^{\mathcal{C}}(\bar{Q}, \bar{Q}_0, g_0)(O_{j2})\Big] \\
\text{with } D^{\mathcal{C}}(\bar{Q}, \bar{Q}_0, g_0)(O_i) &= \left(\frac{\mathbb{I}(A_i = 1)}{g_0(A_i)} - \frac{\mathbb{I}(A_i = 0)}{g_0(A_i)}\right)\big(Y_i - \bar{Q}(A_i, W_i)\big) \\
&\quad - \Big[\big(\bar{Q}_0(1, W_i) - \bar{Q}(1, W_i)\big) - \big(\bar{Q}_0(0, W_i) - \bar{Q}(0, W_i)\big)\Big]
\end{aligned}$$

The second component $\bar{D}^{\mathcal{F}}(\bar{X}_j, \bar{O}_j)$ is the following function of the paired full data $\bar{X}_j = \big(X_{j1}, X_{j2}\big)$:

$$\begin{aligned}
\bar{D}^{\mathcal{F}}(\bar{Q}_0)(\bar{X}_j, \bar{O}_j) &= \frac{1}{2}\Big[D^{\mathcal{F}}(\bar{Q}_0)(X_{j1}, O_{j1}) + D^{\mathcal{F}}(\bar{Q}_0)(X_{j2}, O_{j2})\Big] \\
\text{with } D^{\mathcal{F}}(\bar{Q}_0)(X_i, O_i) &= Y_i(1) - Y_i(0) - \big[\bar{Q}_0(1, W_i) - \bar{Q}_0(0, W_i)\big]
\end{aligned}$$

The standardized TMLE for the SATE is asymptotically normal with mean 0 and variance $\sigma^{2,\mathcal{S}}$ given by the limit of

$$\sigma_n^{2,\mathcal{S}} = \frac{1}{n/2}\sum_{j=1}^{n/2} P_0^n\left\{\bar{D}^{\mathcal{S}}(\bar{Q},\bar{Q}_0,g_0)(\bar{X}_j,\bar{O}_j)\right\}^2$$

where $P_0^n f = E_0[f|W^n]$ denotes the conditional expectation of a function f of $O^n = (O_1,\ldots,O_n)$, given the vector of baseline covariates $W^n = (W_1,\ldots,W_n)$.

Proof. Let $\bar{Q}_0(W) = \bar{Q}_0(1,W) - \bar{Q}_0(0,W)$ denote the true difference in treatment-specific means. We can write the deviation between the TMLE $\Psi_n(P_n)$ for the conditional estimand $\Psi_0^{\mathcal{C}}(P_0)$ and the SATE as

$$\begin{aligned}
&\Psi_n(P_n) - \Psi^{\mathcal{S}}(P_X)\\
&\quad= \Psi_n(P_n) - \Psi_0^{\mathcal{C}}(P_0) - \left[\Psi^{\mathcal{S}}(P_X) - \Psi_0^{\mathcal{C}}(P_0)\right]\\
&\quad= \frac{1}{n/2}\sum_{j=1}^{n/2}\bar{D}^{\mathcal{C}}(\bar{O}_j) - \left[\Psi^{\mathcal{S}}(P_X) - \Psi_0^{\mathcal{C}}(P_0)\right] + o_P(\sqrt{n/2})\\
&\quad= \frac{1}{n/2}\sum_{j=1}^{n/2}\bar{D}^{\mathcal{C}}(\bar{O}_j) - \left[\frac{1}{n}\sum_{i=1}^{n} Y_i(1) - Y_i(0) - \bar{Q}_0(W_i)\right] + o_P(\sqrt{n/2})\\
&\quad= \frac{1}{n/2}\sum_{j=1}^{n/2}\left[\bar{D}^{\mathcal{C}}(\bar{O}_j) - \frac{1}{2}\left(Y_{j1}(1) - Y_{j1}(0) - \bar{Q}_0(W_{j1}) + Y_{j2}(1) - Y_{j2}(0) - \bar{Q}_0(W_{j2})\right)\right] + o_P(\sqrt{n/2})\\
&\quad= \frac{1}{n/2}\sum_{j=1}^{n/2}\left[\bar{D}^{\mathcal{C}}(\bar{O}_j) - \bar{D}^{\mathcal{F}}(\bar{X}_j,\bar{O}_j)\right] + o_P(\sqrt{n/2})
\end{aligned}$$

where $\bar{D}^{\mathcal{C}}(\bar{O}_j)$ is the influence curve of the TMLE for the conditional estimand $\Psi_0^{\mathcal{C}}(P_0)$ under adaptive pair-matching and where $\bar{D}^{\mathcal{F}}(\bar{X}_j,\bar{O}_j)$ is the following function of the paired full data $\bar{X}_j = \big(X_{j1}, X_{j2}\big)$:

$$\bar{D}^{\mathcal{F}}(\bar{X}_j,\bar{O}_j) = \frac{1}{2}\left[D^{\mathcal{F}}(X_{j1},O_{j1}) + D^{\mathcal{F}}(X_{j2},O_{j2})\right]$$
$$\text{with } D^{\mathcal{F}}(X_i,O_i) = Y_i(1) - Y_i(0) - \left[\bar{Q}_0(1,W_i) - \bar{Q}_0(0,W_i)\right]$$

Thus, we have shown the TMLE is an asymptotically linear estimator of the SATE in a trial with adaptive pair-matching:

$$\Psi_n(P_n) - \Psi^{\mathcal{S}}(P_X) = \frac{1}{n/2}\sum_{j=1}^{n/2}\bar{D}^{\mathcal{S}}(\bar{X}_j,\bar{O}_j) + o_P(\sqrt{n/2})$$

with influence curve

$$\bar{D}^{\mathcal{S}}(\bar{X}_j,\bar{O}_j) = \bar{D}^{\mathcal{C}}(\bar{O}_j) - \bar{D}^{\mathcal{F}}(\bar{X}_j,\bar{O}_j).$$

□

Strictly speaking, the influence curve must only be a function of the observed data. Nonetheless, the theorem is sufficient to prove asymptotic normality and consistency of the TMLE.

Appendix B: Variance and variance estimation for the TMLE of the SATE in an Adaptive Pair-Matched Trial

Theorem 2. *The asymptotic variance of the standardized estimator is given by the limit of*

$$\sigma_n^{2,\mathcal{S}} = \frac{1}{n/2}\sum_{j=1}^{n/2} P_0^n\left\{\bar{D}^{\mathcal{S}}(\bar{Q},\bar{Q}_0,g_0)(\bar{X}_j,\bar{O}_j)\right\}^2$$
$$= \frac{1}{n/2}\sum_{j=1}^{n/2}\left[P_0^n\left\{\bar{D}^{\mathcal{C}}(\bar{Q},\bar{Q}_0,g_0)(\bar{O}_j)\right\}^2 - \frac{1}{4}P_0^n\left\{D^{\mathcal{F}}(\bar{Q}_0)(X_{j1},O_{j1})\right\}^2 - \frac{1}{4}P_0^n\left\{D^{\mathcal{F}}(\bar{Q}_0)(X_{j2},O_{j2})\right\}^2\right]$$

Proof. The conditional variance can be expressed as

$$\sigma_n^{2,\mathcal{S}} = \frac{1}{n/2}\sum_{j=1}^{n/2} P_0^n\left\{\bar{D}^{\mathcal{S}}(\bar{X}_j,\bar{O}_j)\right\}^2$$
$$= \frac{1}{n/2}\sum_{j=1}^{n/2} P_0^n\left\{\bar{D}^{\mathcal{C}}(\bar{O}_j) - \bar{D}^{\mathcal{F}}\bar{X}_j,\bar{O}_j)\right\}^2$$
$$= \frac{1}{n/2}\sum_{j=1}^{n/2}\left[P_0^n\left\{\bar{D}^{\mathcal{C}}(\bar{O}_j)\right\}^2 + P_0^n\left\{\bar{D}^{\mathcal{F}}(\bar{X}_j,\bar{O}_j)\right\}^2 - 2P_0^n\left\{\bar{D}^{\mathcal{C}}(\bar{O}_j)\times\bar{D}^{\mathcal{F}}(\bar{X}_j,\bar{O}_j)\right\}\right]$$

The conditional variance of the $\bar{D}^{\mathcal{F}}(\bar{X}_j,\bar{O}_j)$ component is

$$P_0^n\left\{\bar{D}^{\mathcal{F}}(\bar{X}_j,\bar{O}_j)\right\}^2 = P_0^n\left\{\frac{1}{2}\big(D^{\mathcal{F}}(X_{j1},O_{j1}) + D^{\mathcal{F}}(X_{j2},O_{j2})\big)\right\}^2$$
$$= \frac{1}{4}P_0^n\left\{D^{\mathcal{F}}(X_{j1},O_{j1})\right\}^2 + \frac{1}{4}P_0^n\left\{D^{\mathcal{F}}(X_{j2},O_{j2})\right\}^2$$
$$+ \frac{1}{2}P_0^n\left\{D^{\mathcal{F}}(X_{j1},O_{j1})\times D^{\mathcal{F}}(X_{j2},O_{j2})\right\}$$

Under the randomization assumption, each $D^{\mathcal{F}}(X_i,O_i)$ component has conditional mean zero, given its baseline covariates W_i:

$$E[D^{\mathcal{F}}(X_i,O_i)\big|W_i] = E\big[Y_i(1) - Y_i(0) - \bar{Q}_0(W_i)|W_i\big]$$
$$= E[Y_i(1)|W_i] - E[Y_i(0)|W_i] - \big[\bar{Q}_0(1,W_i) - \bar{Q}_0(0,W_i)\big]$$
$$= E_0(Y_i|A_i=1,W_i) - E_0(Y_i|A_i=0,W_i) - \big[\bar{Q}_0(1,W_i) - \bar{Q}_0(0,W_i)\big] = 0$$

Therefore, the conditional covariance of the $D^{\mathcal{F}}$ components within a matched pair is zero:

$$P_0^n\left\{D^{\mathcal{F}}(X_{j1},O_{j1})\times D^{\mathcal{F}}(X_{j2},O_{j2})\right\} = E\big[D^{\mathcal{F}}(X_{j1},O_{j1})\times D^{\mathcal{F}}(X_{j2},O_{j2})\big|W^n\big]$$
$$= E[E\big[Y_{j1}(1) - Y_{j1}(0) - \bar{Q}_0(W_{j1})|W_{j1}\big]\times D^{\mathcal{F}}(X_{j2},O_{j2})] = 0$$

The conditional variance of the $\bar{D}^{\mathcal{F}}(\bar{X}_j,\bar{O}_j)$ component simplifies to

$$P_0^n\left\{\bar{D}^{\mathcal{F}}(\bar{X}_j,\bar{O}_j)\right\}^2 = \frac{1}{4}P_0^n\left\{D^{\mathcal{F}}(X_{j1},O_{j1})\right\}^2 + \frac{1}{4}P_0^n\left\{D^{\mathcal{F}}(X_{j2},O_{j2})\right\}^2$$

The conditional covariance of the $\bar{D}^{\mathcal{C}}(\bar{O}_j)$ and $\bar{D}^{\mathcal{F}}(\bar{X}_j, \bar{O}_j)$ components is

$$\begin{aligned} P_0^n\left\{\bar{D}^{\mathcal{C}}(\bar{O}_j) \times \bar{D}^{\mathcal{F}}(\bar{X}_j, \bar{O}_j)\right\} &= \frac{1}{4} P_0^n\left\{\left[D^{\mathcal{C}}(O_{j1}) + D^{\mathcal{C}}(O_{j2})\right] \times \left[D^{\mathcal{F}}(X_{j1}, O_{j1}) + D^{\mathcal{F}}(X_{j2}, O_{j2})\right]\right\} \\ &= \frac{1}{4}\Big[P_0^n\left\{D^{\mathcal{C}}(O_{j1}) \times D^{\mathcal{F}}(X_{j1}, O_{j1})\right\} + P_0^n\left\{D^{\mathcal{C}}(O_{j1}) \times D^{\mathcal{F}}(X_{j2}, O_{j2})\right\} \\ &\quad + P_0^n\left\{D^{\mathcal{C}}(O_{j2}) \times D^{\mathcal{F}}(X_{j1}, O_{j1})\right\} + P_0^n\left\{D^{\mathcal{C}}(O_{j2}) \times D^{\mathcal{F}}(X_{j2}, O_{j2})\right\}\Big] \end{aligned}$$

As shown in Appendix of Balzer et al. (2015a), the covariance of the $D^{\mathcal{C}}(O_i)$ and $D^{\mathcal{F}}(O_i)$ components is equal to the variance of $D^{\mathcal{F}}(O_i)$. Therefore, we have

$$\begin{aligned} P_0^n\left\{D^{\mathcal{C}}(O_{j1}) \times D^{\mathcal{F}}(X_{j1}, O_{j1})\right\} &= P_0^n\left\{D^{\mathcal{F}}(X_{j1}, O_{j2})\right\}^2 \\ P_0^n\left\{D^{\mathcal{C}}(O_{j2}) \times D^{\mathcal{F}}(X_{j2}, O_{j2})\right\} &= P_0^n\left\{D^{\mathcal{F}}(X_{j2}, O_{j2})\right\}^2 \end{aligned}$$

Under the randomization assumption, the other terms are zero:

$$\begin{aligned} P_0^n\left\{D^{\mathcal{C}}(O_{j1}) \times D^{\mathcal{F}}(X_{j2}, O_{j2})\right\} &= E\left[D^{\mathcal{C}}(O_{j1}) \times E\left[D^{\mathcal{F}}(X_{j2}, O_{j2})\middle|W_{j2}\right]\right] = 0 \\ P_0^n\left\{D^{\mathcal{C}}(O_{j2}) \times D^{\mathcal{F}}(X_{j1}, O_{j1})\right\} &= E\left[D^{\mathcal{C}}(O_{j2}) \times E\left[D^{\mathcal{F}}(X_{j1}, O_{j1})\middle|W_{j1}\right]\right] = 0 \end{aligned}$$

We have that the conditional covariance of the $\bar{D}^{\mathcal{C}}(\bar{O}_j)$ and $\bar{D}^{\mathcal{F}}(\bar{X}_j, \bar{O}_j)$ components equals

$$P_0^n\left\{\bar{D}^{\mathcal{C}}(\bar{O}_j) \times \bar{D}^{\mathcal{F}}(\bar{X}_j, \bar{O}_j)\right\} = \frac{1}{4} P_0^n\left\{D^{\mathcal{F}}(X_{j1}, O_{j2})\right\}^2 + \frac{1}{4} P_0^n\left\{D^{\mathcal{F}}(X_{j2}, O_{j2})\right\}^2$$

Combining the terms, we have

$$\sigma_n^{2,\mathcal{S}} = \frac{1}{n/2} \sum_{j=1}^{n/2} \left[P_0^n\left\{\bar{D}^{\mathcal{C}}(\bar{O}_j)\right\}^2 - \frac{1}{4} P_0^n\left\{D^{\mathcal{F}}(X_{j1}, O_{j1})\right\}^2 - \frac{1}{4} P_0^n\left\{D^{\mathcal{F}}(X_{j2}, O_{j2})\right\}^2\right]$$

□

The asymptotic variance of the TMLE for the SATE $\sigma^{2,\mathcal{S}}$ is always less than or equal to $\sigma^{2,\mathcal{C}}$, which is the asymptotic variance of the TMLE for the conditional parameter. As shown in Balzer et al. (2015b), we can estimate the upper bound as

$$\begin{aligned} \hat{\sigma}^{2,\mathcal{S}} = \hat{\sigma}^{2,\mathcal{C}} &= \frac{1}{n/2} \sum_{j=1}^{n/2} \left\{\hat{\bar{D}}^{\mathcal{C}}(\bar{Q}_n^*, g_0)(\bar{O}_j)\right\}^2 \\ \hat{\bar{D}}^{\mathcal{C}}(\bar{Q}_n^*, g_0)(\bar{O}_j) &= \frac{1}{2}\left[\hat{D}^{\mathcal{C}}(\bar{Q}_n^*, g_0)(O_{j1}) + \hat{D}^{\mathcal{C}}(\bar{Q}_n^*, g_0)(O_{j2})\right] \\ \hat{D}^{\mathcal{C}}(\bar{Q}_n^*, g_0)(O_i) &= \left(\frac{\mathbb{I}(A_i = 1)}{g_0(A_i)} - \frac{\mathbb{I}(A_i = 0)}{g_0(A_i)}\right)\left(Y_i - \bar{Q}_n^*(A_i, W_i)\right) \end{aligned}$$

Ordering the observations within matched pairs, such that the first corresponds to the unit randomized to the intervention ($A_{j1} = 1$) and the second to the control ($A_{j2} = 0$), it follows that

$$\hat{\bar{D}}^{\mathcal{C}}(\bar{Q}_n^*, g_0)(\bar{O}_j) = \hat{\bar{D}}^{\mathcal{C}}(\bar{Q}_n^*, g_0)(\bar{O}_j) = \big(Y_{j1} - \bar{Q}_n^*(1, W_{j1})\big) - \big(Y_{j2} - \bar{Q}_n^*(0, W_{j2})\big)$$

allowing us to represent the variance estimator as the sample variance of the difference in residuals within matched pairs:

$$\hat{\sigma}^{2,\mathcal{S}} = \hat{\sigma}^{2,\mathcal{C}} = \frac{1}{n/2} \sum_{j=1}^{n/2} \left\{ \big(Y_{j1} - \bar{Q}_n^*(1, W_{j1})\big) - \big(Y_{j2} - \bar{Q}_n^*(0, W_{j2})\big) \right\}^2$$

This variance estimator will be consistent if there is no heterogeneity in the treatment effect within strata of covariates (i.e. if the variance of the $D^{\mathcal{F}}$ component is zero) *and* if the conditional mean function $\bar{Q}_0(A, W)$ is consistently estimated. Otherwise, the variance estimator will be conservative.

We can relax the assumption that the conditional variance converges to some limit. Specifically, we have that the standardized and scaled estimator converges to a normal distribution with mean 0 and variance given by the ratio of the true conditional variance $\sigma_n^{2,\mathcal{S}}$ divided by our conservative estimator $\hat{\sigma}^{2,\mathcal{S}}$:

$$\frac{\Psi(P_n) - \Psi^{\mathcal{S}}(P_X)}{\hat{\sigma}^{2,\mathcal{S}}/\sqrt{n/2}} \to N\left(0, \frac{\sigma_n^{2,\mathcal{S}}}{\hat{\sigma}^{2,\mathcal{S}}}\right)$$

Since the ratio of variances is always ≤ 1. The standard normal distribution $N(0, 1)$ provides a conservative approximation to the asymptotic distribution.

Appendix C: Step-by-step instructions to obtain a cross-validated variance estimator

Let $\bar{Q}_n(A, W)$ denote the initial estimator for conditional mean outcome, which was selected through the data-adaptive procedure (Sec. 3.1 for a non-matched trial and Sec. 4.1 for a pair-matched trial). Let $g_n(A|W)$ denote the estimator of the exposure mechanism, which was collaboratively selected through the data-adaptive procedure (Sec. 5.1). A cross-validated estimate of the variance of the data-adaptive TMLE can be implemented as follows. As before, we present V-fold cross-validation, where the data are partitioned into V folds of size $\approx n/V$. If matching was used, the partitioning should preserve the pairs.

i. For each fold $v = \{1, \ldots, V\}$ in turn,

   a. Set the observation(s) in fold v to be the validation set and the remaining observations to be the training set.

   b. Using observations in the training set, fit the selected TMLE.
      - Fit the selected working model for the conditional mean outcome $\bar{Q}_n(A, W)$.
      - Fit the selected working model for the exposure mechanism $g_n(A|W)$.
      - Target the initial estimator. Denote the estimated fluctuation coefficient ϵ_n.

   c. For each observation O_k in the validation set, estimate the influence curve (and correction factor ρ_n if relevant).
      - Use the initial fit $\bar{Q}_n(A, W)$, based on the training data, to obtain initial predictions of the outcome under the treatment $\bar{Q}_n(1, W_k)$ and under the control $\bar{Q}_n(0, W_k)$.

- Use the the fit of the exposure mechanism $g_n(A, W)$, based on the training set, to calculate the clever covariate $H_n(A_k, W_k)$.
- Update the initial estimates with the estimated fluctuation parameter ϵ_n. Denote the targeted predictions of the outcome under the treatment $\bar{Q}^*_n(1, W_k)$ and the control $\bar{Q}^*_n(0, W_k)$.
- Plug-in the relevant components to estimate influence curve, appropriate for the target parameter and study design.

ii. Estimate the variance of the data-adaptive TMLE with the sample variance of the estimated influence curve, normalized by the appropriate sample size.

Appendix D: Sample R Code

```
> ####################################
> # Simulations to illustrate estimation and inference for the PATE
> #  and SATE with the unadjusted estimator, the MLE with a priori specified
> #        adjustment set (W.9), TMLE with adaptive pre-specification for initial
> # estimation of Qbar(A,W) and C-TMLE including collaborative estimation of g(A|W)
> #
> # Programmer: Laura Balzer (lbbalzer@berkeley.edu)
> #
> # Last update: 05.13.15
> ####################################
>
>
> #-----------------------------------
> # simulate.Data.and.Run: function to generate the simulated data
> #        and run the estimators
> #
> # input:
> # output: true values of PATE and SATE, point estimates and inference from the 4
> # estimators, results from CV-selection
> #-----------------------------------
>
> simulate.Data.and.Run<- function(){
+
+   X<- get.Data(n)
+
+   # PATE is the average difference in counterfactuals over the population
+   PATE<<- get.PATE()
+
+   # SATE is the average of the difference in counterfactuals for the sample
+   SATE<- mean(X$Y1- X$Y0)
+
+   # True value of the target parameter depending on the scientific question
+   truth<- ifelse(POP.EFFECT, PATE, SATE)
+
+   # If Matching, use the nbpMatching package function to create pairs
```

```
+   if(MATCHING){
+
+     # see nbpMatching package for further details
+     dist<- distancematrix(gendistance(data.frame(X[, matchOn])))
+     matches<- nonbimatch(dist)
+     # matches contains ids for the pair match as well as the distance measure
+     grpA<- as.numeric(matches$halves[,'Group1.Row'])
+     grpB<- as.numeric(matches$halves[,'Group2.Row'])
+
+     # The paired data structure is organized such that the first observation in
+     #pair j is on row j, while the second obs. in pair j is on row (j+n.pairs)
+
+     X.all<- rbind( X[grpA, ], X[grpB, ] )
+     # first obs in the matched pair us on rows 1:n.pairs
+     # second obs in the matched pair is on rows (n.pairs: n)
+
+   }        else{ # no matching
+     X.all<- X
+   }
+
+
+   #---------------------
+   # Assign intervention and observed data
+   #---------------------
+
+   # randomly assign treatment so that it's balanced overall
+   A<- rbinom(n.pairs, 1, 0.5)
+   A.2<- ifelse(A==1, 0, 1)
+   A <- c(A,A.2)
+
+   # we observe the counterfactual outcome corresponding to
+   #        the observed exposure
+   Y<- ifelse(A, X.all$Y1, X.all$Y0)
+
+   # observed data are (W, A, Y)
+   data<- data.frame(X.all[,1:12], A, Y)
+
+   #--------------------------
+   # ESTIMATION AND INFERENCE
+   #--------------------------
+
+   # unadjusted estimator
+   unadj<- do.Estimation(Do.CV.Inference=F, truth=truth, data=data, QAdj='U',
+                         family=FAMILY)
+
+   # a priori spec. working model for Qbar(A,W), adjusting for W9
+   mle.W9<- do.Estimation(Do.CV.Inference=F, truth=truth, data=data, QAdj='W.9',
+                          family=FAMILY)
```

```
+
+   # Adaptive Pre-Specified Approach for Step 1. Initial Estimation
+   select.Q<- suppressWarnings( CV.selector(data, family=FAMILY, forQ=T,
+                                            gAdj=NULL))
+
+   # TMLE based on this selection (need to do CV-inference)
+   tmle<- do.Estimation(Do.CV.Inference=T, truth=truth, data=data, QAdj=select.Q,
+                        family=FAMILY)
+
+   # Adaptive Pre-Specified Approach for Step 2. Targeting
+   select.G<- CV.selector(data,  family=FAMILY, forQ=F, QAdj=select.Q)
+
+   # C-TMLE using the adaptive selection for Step1 and Step2
+   ctmle<-  do.Estimation(Do.CV.Inference=T, truth=truth, data=data,
+                          QAdj=select.Q, family=FAMILY, gAdj=select.G)
+
+   CV.select<- data.frame(select.Q, select.G)
+
+   RETURN<- list(PATE=PATE, SATE=SATE, unadj=unadj, mle.W9=mle.W9, tmle=tmle,
+                 ctmle=ctmle, CV.select=CV.select)
+
+   RETURN
+ }
> #------------------
> # do.Estimation: function to do TMLE + get inference
> #
> # input: Do.CV.Inference (get cross-validated estimate of IC),
> #   truth (true value of the target parameter), obs. data,
> #   QAdj (adjustment variable for Qbar(A,W)), family for Qbar(A,W),
> #                gAdj (adjustment variable for g(A|W)), significance level (alpha)
> #
> # output: point estimate, inference, CV-inference (if appropriate)
> #-----------------
>
> do.Estimation<- function(Do.CV.Inference=F, truth, data, QAdj, family,
+                          gAdj=NULL, alpha=0.05){
+
+   # get a point estimate with TMLE based on the full data (i.e. without CV)
+   est<- suppressWarnings(do.TMLE.with.CV(Do.CV=F, train=data, QAdj=QAdj,
+                                          family=family, gAdj=gAdj))
+
+   # estimate the influence curve (and residuals)
+   resid<- (data$Y - est$QbarAW)
+   DY<- est$H.AW*(data$Y - est$QbarAW)
+   DW<- est$Qbar1W - est$Qbar0W - est$psi.hat
+
+   # get the variance of the influence curve (appropriate for the study design)
+   var.IC<- get.Variance.IC(resid=resid, DY=DY, DW=DW)
```

```
+
+   # degrees of freedom for t-distribution
+   df<- ifelse(PAIRED, (n.pairs-1),  (n-2) )
+
+   # get confidence interval coverage and rejection
+   if(POP.EFFECT){
+     inference<-get.Inference.tdist(truth=truth, psi.hat=est$psi.hat, alpha=alpha,
+                                    df=df, var= var.IC$var.PATE)
+   } else{
+     inference<-get.Inference.tdist(truth=truth, psi.hat=est$psi.hat, alpha=alpha,
+                                    df=df, var= var.IC$var.SATE)
+   }
+
+   #-------------------------------------
+   # IF GETTING A CROSS-VALIDATED ESTIMATE OF THE INFLUENCE CURVE
+   #                (Section 6 of manuscript)
+   #-------------------------------------
+   if(Do.CV.Inference){
+
+     IC.CV<- get.CV.inference(data=data, QAdj=QAdj, family=family, gAdj=gAdj)
+     var.IC.CV<- get.Variance.IC(resid=IC.CV$resid, DY=IC.CV$DY, DW=IC.CV$DW)
+
+     if(POP.EFFECT){
+       inference.CV<- get.Inference.tdist(truth=truth, psi.hat=est$psi.hat,
+                                          alpha=alpha, df=df, var= var.IC.CV$var.PATE)
+     }else{
+       inference.CV<- get.Inference.tdist(truth=truth, psi.hat=est$psi.hat,
+                                          alpha=alpha, df=df, var= var.IC.CV$var.SATE)
+     }
+
+   } else{ # otherwise, fill in the CV-inference with NA
+     inference.CV<- data.frame(var=NA, tstat=NA, cover=NA, reject=NA)
+   }
+   #-------------------------------------
+
+   RETURN<- data.frame(est$psi.hat,         inference, inference.CV)
+
+   colnames(RETURN)<- c('psi.hat','var','tstat','cover','reject', 'var.cv',
+                        'tstat.cv', 'cover.cv', 'reject.cv')
+   RETURN
+ }
> #--------------------------
> # do.TMLE.with.CV: runs the standard TMLE algorithm (i.e. training set= data);
> #   will also fit the TMLE algorithm for Qbar*(A,W) on training set
> #   and get estimates for validation set.
> #
> #      input: Do.CV (whether not doing cross-validation), train (training set),
> #   valid (validation set), Qadj (adjustment variable for Qbar(A,W)),
```

```
> #   family for fitting Qbar(A,W)), gAdj (adjustment variable for g(A|W))
> #
> # output: estimates of the propensity score (exposure mechanism),
> #   clever covariate (H.AW), targeted predictions on obs exposure, txt & control,
> #   and point estimate (if not doing cross-validation)
> #----------------------
>
> do.TMLE.with.CV<- function(Do.CV=F, train, valid, QAdj, family, gAdj=NULL){
+
+   # do entire TMLE algorithm on the training set
+   train.temp<- train[, c(QAdj, 'A', 'Y')]
+
+   # STEP 1: INITIAL ESTIMATION
+   # fit the working model for Qbar(A,W) on the training set
+   glm.out<-  glm( Y~., family=family,  data=train.temp )
+
+   # get the predicted outcomes under obs exp, txt and control
+   X1<- X0<- train
+   X1$A<-1; X0$A<- 0
+
+   QbarAW.train <- predict(glm.out, newdata=train, type='response')
+   Qbar1W.train<-  predict(glm.out, newdata=X1, type='response')
+   Qbar0W.train<-  predict(glm.out, newdata=X0, type='response')
+
+   #-------------------
+   # STEP 2: TARGETING
+   # estimate the propensity score (i.e. the exposure mechanism)
+
+   if( is.null(gAdj) ){
+     # if adj var for g(A|W) is null, then g(A|W)=0.5
+     pscore.train <-  rep(0.5, nrow(train) )
+   } else if (gAdj=='U'){
+     # if adj var for g(A|W) is 'U' (unadj), then g(A|W)=0.5
+     pscore.train <-  rep(0.5, nrow(train) )
+   } else{
+     # otherwise fit the working model for the pscore on training set
+     train.temp<- train[, c(gAdj, 'A')]
+     p.out<-   glm( A~., family='binomial', data=train.temp)
+     pscore.train <- predict(p.out, newdata=train,  type="response")
+   }
+
+   # calculate the clever covariate
+   H.AW.train<- train$A/pscore.train - (1-train$A)/(1-pscore.train)
+
+   # update the initial estimator of the outcome regression
+
+   if(family!='binomial'){
+     # if cont. and unbounded outcome, linear update
```

```
+     linUpdate<-  glm(train$Y ~ -1 +offset( QbarAW.train) + H.AW.train,
+                      family="gaussian")
+     eps<-linUpdate$coef
+
+     # updated QbarAW estimates for training set.
+     QbarAW.train<- QbarAW.train + eps*H.AW.train
+     Qbar1W.train<- Qbar1W.train + eps/pscore.train
+     Qbar0W.train<- Qbar0W.train - eps/(1-pscore.train)
+
+   }else { # if binary or bounded outcome in [0,1], logistic update
+
+     logitUpdate<- suppressWarnings( glm(train$Y ~ -1 +offset(qlogis(QbarAW.train))
+                                         + H.AW.train, family="binomial"))
+     eps<-logitUpdate$coef
+
+     # updated QbarAW estimates for training set.
+     QbarAW.train<- plogis( qlogis(QbarAW.train) + eps*H.AW.train)
+     Qbar1W.train<- plogis( qlogis(Qbar1W.train) + eps/pscore.train)
+     Qbar0W.train<- plogis( qlogis(Qbar0W.train) - eps/(1-pscore.train))
+   }
+
+   #-----------------------
+   # if not doing cross-validation (i.e. if running the std TMLE algorithm)
+   if(!Do.CV){
+
+     # STEP 3: PARAMETER ESTIMATION
+     psi.hat<- mean(Qbar1W.train- Qbar0W.train)
+
+     RETURN<- list(pscore=pscore.train, H.AW=H.AW.train, QbarAW=QbarAW.train,
+                   Qbar1W=Qbar1W.train, Qbar0W=Qbar0W.train, psi.hat=psi.hat)
+
+   } else{
+     # If the goal is to obtain a Cross-validated variance estimator, get the
+     # initial estimates and updates for the validation set
+
+     # get initial estimates based on the fit of Qbar(A,W) from the training set
+     V1<- V0<- valid
+     V1$A= 1; V0$A=0
+
+     QbarAW.valid<- predict(glm.out, newdata=valid, type='response')
+     Qbar1W.valid<- predict(glm.out, newdata=V1, type='response')
+     Qbar0W.valid<- predict(glm.out, newdata=V0, type='response')
+
+     # get estimates of the propensity score based on the fit of g(A|W)
+     #       from the training set
+     if( is.null(gAdj)){
+       pscore.valid<- rep(0.5, nrow(valid))
+     } else if (gAdj=='U'){
```

```
+       pscore.valid<- rep(0.5, nrow(valid))
+     } else{
+       pscore.valid<- predict(p.out, newdata=valid,  type='response')
+     }
+
+     # calculate the clever covariate
+     H.AW.valid<- valid$A/pscore.valid - (1-valid$A)/(1-pscore.valid)
+
+     # update
+     if(family!='binomial'){
+       QbarAW.valid<- QbarAW.valid + eps*H.AW.valid
+       Qbar1W.valid<- Qbar1W.valid + eps/pscore.valid
+       QbarOW.valid<- QbarOW.valid - eps/(1-pscore.valid)
+     } else {
+       QbarAW.valid<- plogis( qlogis(QbarAW.valid) + eps*H.AW.valid)
+       Qbar1W.valid<- plogis( qlogis(Qbar1W.valid) + eps/pscore.valid)
+       QbarOW.valid<- plogis( qlogis(QbarOW.valid) - eps/(1-pscore.valid))
+     }
+
+     RETURN<- list(pscore=pscore.valid, H.AW=H.AW.valid, QbarAW=QbarAW.valid,
+                   Qbar1W=Qbar1W.valid, QbarOW=QbarOW.valid)
+   }
+   RETURN
+ }
> #------------
> # get.Variance.IC: function to get the variance based on the estimated influence
> # curve (IC)
> #         input: residuals + estimates of the DY and DW components of the IC
> #          output: variance for PATE or SATE depending on the study design
> #-------------
>
> get.Variance.IC<- function(resid, DY, DW){
+
+   # without pair-matching
+   if(!PAIRED){
+     var.PATE<- (var(DY + DW) )/n         # (Eq. 3)
+     var.SATE<- var(DY)/n  # (Eq. 4)
+
+   } else       { # if pair-matched trial
+
+     # note data are organized such that first units in matched-pairs are
+     # (1:n.pairs) and second units are (n.pairs+1 to n)
+
+     # var for the PATE approx with 1/n x (sample var if non-paired IC - 2
+     # covariance of the residuals) (section 4)
+     rho <- cov(resid[1:n.pairs], resid[(n.pairs+1):n])  #(eq. 5)
+     var.PATE<- (var(DY + DW) - 2*rho)/n
+
```

```
+     # DbarY= 1/2 sum_{i in pairs} HAW_i*(Y_i -Qbar_i)
+     # Serves as the upper bound on the IC for SATE
+     DY.paired <-   .5*(DY[1:n.pairs] + DY[(n.pairs+1):n])                #(eq. 6)
+     var.SATE<- var(DY.paired)/n.pairs
+   }
+
+   data.frame(var.PATE, var.SATE)
+ }
> #---------------------
> # get.CV.inference: function to get an CV-estimate of the IC
> #        also used by the CV-selector to choose the candidate with the smallest CV-risk
> #        (i.e. smallest estimated variance based on CV-estimate of the IC)
> #
> #        input: data, QAdj (adjustment variable for Qbar(A,W) ), family for Qbar(A,W),
> #                gAdj (adjustment variable for g(A|W) )
> #
> # output: cross-validated estimates of the clever covariate H.AW,
> #        the residuals (Y-Qbar*(A,W)), DY component of the IC and
> #         DW component of the IC
> #--------------
>
> get.CV.inference<- function(data, QAdj, family, gAdj){
+
+   H.AW<- QbarAW <- Qbar1W <- Qbar0W<- rep(NA,n)
+
+   # run the TMLE algorithm using data in the training set
+   # and obtain estimates of the clever covariate H.AW and targeted predictions
+   # for observations Qbar*(A,W), Qbar*(1,W), Qbar*(0,W) in the validation set.
+   for(i in 1: nFolds) {
+
+     valid <- data[fold==i, ]
+     train <- data[fold!=i,]
+
+     temp<- do.TMLE.with.CV(Do.CV=T, train=train, valid=valid, QAdj=QAdj,
+                            family=family,  gAdj=gAdj)
+
+     H.AW[fold==i]<- temp$H.AW
+     QbarAW[fold==i]<- temp$QbarAW
+     Qbar1W[fold==i]<- temp$Qbar1W
+     Qbar0W[fold==i]<- temp$Qbar0W
+   }
+
+   # calculate the relevant components of the influence curve (and the residuals)
+   psi.hat<- mean(Qbar1W - Qbar0W)
+
+   resid <- data$Y - QbarAW
+   DY <- H.AW*(data$Y - QbarAW)
+   DW <-  Qbar1W - Qbar0W- psi.hat
```

```
+
+   data.frame(H.AW, resid, DY, DW)
+ }
> #----------------------------
> # get.Inference.tdist: get inference assuming a Student's t-distribution
> #
> #        input: truth (true value of psi), psi.hat (point estimate), alpha
> # (signifance level), df (degrees of freedom) var (variance estimate)
> #
> # output: variance, tstat, coverage, reject null.
> #--------
>
> get.Inference.tdist<- function(truth, psi.hat, alpha, df, var){
+
+   # cutoff based on t-dist for testing and CI
+   cutoff <- qt(alpha/2, df=df, lower.tail=F)
+
+   # standard error (square root of the variance)
+   se<- sqrt(var)
+
+   # confidence interval coverage
+   cover<- ( (psi.hat - cutoff*se) <= truth & truth <= (psi.hat + cutoff*se) )
+
+   # test statistic and pvalue
+   tstat <- psi.hat/se
+   reject<- 2*pt(abs(tstat), df=df, lower.tail=F) < alpha
+
+   data.frame(var, tstat, cover, reject)
+ }
> #-----------------------------------------
> # CV.selector: function to select the candidate working model (TMLE)
> # that minimizes the estimated variance for Step 1 Initial estimation
> #        and in Step 2 Targeting.
> #
> # input: data, family for Qbar(A,W), forQ (CV selection for Qbar(A,W) or g(A|W))
> #          QAdj (a priori spec adj variable for Qbar(A,W)),
> #   gAdj (a priori spec adj variable for g(A|W))
> #
> #        output: candidate adjustment variable with smallest CV-risk
> #-----------------------------------------
>
> CV.selector<- function(data, family, forQ, QAdj, gAdj){
+
+   # number of potential TMLEs correspond to num of possible adj variables
+   num.tmles <- length(ADJ.SET)
+   CV.risk <- rep(NA, num.tmles)
+
+   # since the folds are initialized before running any algorithms,
```

```
+   # we are going to loop through the candidate estimators
+   for(k in 1:num.tmles){
+
+     if(forQ){
+       # adaptive pre-specified approach for Step 1: Initial Estimation
+       # (Sec. 3.1 for non-matched trial and Sec. 4.1 for matched trial)
+       IC.temp<- get.CV.inference(data=data, QAdj=ADJ.SET[k], family=family,
+                                  gAdj=NULL)
+
+     } else{
+       # adaptive pre-specified approach for Step 2: Targeting (based on selected
+       # adjustment variable for Qbar(A,W)). (Sec 5.1)
+       IC.temp<- get.CV.inference(data=data, QAdj=QAdj, family=family,
+                                  gAdj=ADJ.SET[k])
+
+     }
+
+     # CALCULATE THE CV-RISK AS THE CV-VARIANCE ESTIMATE
+     #        APPR. FOR STUDY DESIGN + TARGET PARAMETER
+
+     if(!PAIRED){
+       if(POP.EFFECT){
+         CV.risk[k]<- var(IC.temp$DY +IC.temp$DW)
+       }else{
+         CV.risk[k]<- var(IC.temp$DY)
+       }
+
+     }else{
+       if(POP.EFFECT){
+         rho.temp<-  2*cov(IC.temp$resid[1:n.pairs], IC.temp$resid[(n.pairs+1):n])
+         CV.risk[k]<- var(IC.temp$DY +IC.temp$DW) - rho.temp
+       } else{
+         CV.risk[k]<- var(.5*(IC.temp$DY[1:n.pairs] + IC.temp$DY[(n.pairs+1):n]))
+       }
+     }
+   }
+
+   # SELECT THE ADJUSTMENT VARIABLE RESULTING IN THE SMALLEST
+   # CV-RISK ESTIMATE
+   ADJ.SET[ which.min(CV.risk)]
+ }
> ######################################
> # functions to generate the simulated data (Simulation Study 1)
> # and calculate the true value of the Population Average Treatment Effect (PATE)
> ######################################
>
> get.Data<- function(n){
+
```

```
+   # 9 baseline covariates with specified correlation structure
+   s<- 1
+   Sigma<- matrix(0.5*s*s, nrow=3, ncol=3)
+   diag(Sigma)<- s^2
+
+   W1<- mvrnorm(n, rep(0,3), Sigma)
+   W2<- mvrnorm(n, rep(0,3), Sigma)
+   W3<- cbind(rnorm(n,0,s), rnorm(n,0,s), rnorm(n,0,s))
+
+   # unmeasured factor impacting the outcome
+   UY<- rnorm(n,0,1)
+
+   # calculate the counterfactuals
+   Y0<- get.Y(W1=W1, W2=W2, W3=W3, UY=UY, A=0)
+   Y1<- get.Y(W1=W1, W2=W2, W3=W3, UY=UY, A=1)
+
+   # add-on a dummy variable and return
+   data.frame(U=rep(1,n), W=cbind(W1,W2,W3), Y1=Y1, Y0=Y0)
+ }
> get.Y<- function(W1, W2, W3, UY, A){
+   0.4*A+.25*W1[,1]+.25*W1[,2]+.25*W2[,1]+.25*W2[,2]+.25*UY+.25*A*W1[,1]+.25*A*UY
+ }
> # calculate the population average treatment effect
> #  as average diff in counterfactuals over a pop of 900,000
> get.PATE<- function(N=900000){
+
+   X<- get.Data(n=N)
+   mean(X$Y1 - X$Y0)
+ }
> ##########################
> #---------------------------------
> ##########################
>
> library('nbpMatching')
> library('MASS')
> set.seed(1)
> # Specify sample size
> n<<- 40
> n.pairs<<- n/2
> # Specify the target of inference as Population ATE or Sample ATE
> POP.EFFECT<<- F
> # Specify the study design
> MATCHING <<- T
> # Specify the potential adjustment set (i.e. the library)
> ADJ.SET<<- c('U',paste('W', 1:9, sep='.'))
> FAMILY<<- 'gaussian' # for the conditional mean outcome Qbar(A,W)
> # Our library for initial estimation of Qbar(A,W) consists of linear working
> # regression models  with a main term for the exposure and one candidate
```

```
> # adjustment variable: e.g. Qbar(A,W) = beta0 + beta1*A + beta2*W.1
>
> # Our library for collaborative estimation of g(A|W) consists of logistic working
> # regression models        with a main term for one candidate adjustment variable
> # e.g. logit[g(A|W)]= beta0 + beta1*W.1
>
>
> #---------------------------------
> # Set up the folds for the cross-validation scheme
> if(!MATCHING){ # if it's a non-matched trial
+
+   PAIRED<<- F
+   # do leave-one-out cross-validation
+   fold<- 1:n
+   nFolds<<- n
+
+ } else{
+   PAIRED<<- T
+
+   # The paired data structure is organized such that the first observation in
+   #        pair j is on row j, while the second obs. in pair j is on row (j+n.pairs)
+   fold<<- c( 1:n.pairs, 1:n.pairs)
+   nFolds<<- n.pairs
+
+   # also specify the matching set (covariates W1...W6)
+   matchOn<<- c('W.1', 'W.2', 'W.3', 'W.4', 'W.5', 'W.6')
+ }
> #---------------------------------
>
>
> simulate.Data.and.Run()

[1] "Note: Distances scaled by 0.01 to ensure all data can be handled"
$PATE
[1] 0.3996057

$SATE
[1] 0.3894403

$unadj
    psi.hat        var    tstat cover reject var.cv tstat.cv cover.cv reject.cv
1 0.3767031 0.03244586 2.091315  TRUE  FALSE     NA       NA       NA        NA

$mle.W9
    psi.hat        var    tstat cover reject var.cv tstat.cv cover.cv reject.cv
1 0.3716297 0.03309701 2.042754  TRUE  FALSE     NA       NA       NA        NA

$tmle
```

```
    psi.hat        var    tstat cover reject     var.cv tstat.cv cover.cv
1 0.4149864 0.02295106 2.739255  TRUE   TRUE 0.02628416 2.559686     TRUE
  reject.cv
1      TRUE

$ctmle
    psi.hat        var    tstat cover reject     var.cv tstat.cv cover.cv
1 0.3696772 0.02240197 2.469901  TRUE   TRUE 0.02621188 2.283356     TRUE
  reject.cv
1      TRUE

$CV.select
  select.Q select.G
1      W.5      W.1

>
>
```