Research Article

Statistics in Medicine

Received 13 April 2015, Accepted 20 March 2016 Published online 18 April 2016 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6965

Targeted estimation and inference for the sample average treatment effect in trials with and without pair-matching

Laura B. Balzer,[a*†] Maya L. Petersen,[b] Mark J. van der Laan[b] and the SEARCH Collaboration[a]

In cluster randomized trials, the study units usually are not a simple random sample from some clearly defined target population. Instead, the target population tends to be hypothetical or ill-defined, and the selection of study units tends to be systematic, driven by logistical and practical considerations. As a result, the population average treatment effect (PATE) may be neither well defined nor easily interpretable. In contrast, the sample average treatment effect (SATE) is the mean difference in the counterfactual outcomes for the study units. The sample parameter is easily interpretable and arguably the most relevant when the study units are not sampled from some specific super-population of interest. Furthermore, in most settings, the sample parameter will be estimated more efficiently than the population parameter. To the best of our knowledge, this is the first paper to propose using targeted maximum likelihood estimation (TMLE) for estimation and inference of the sample effect in trials with and without pair-matching. We study the asymptotic and finite sample properties of the TMLE for the sample effect and provide a conservative variance estimator. Finite sample simulations illustrate the potential gains in precision and power from selecting the sample effect as the target of inference. This work is motivated by the Sustainable East Africa Research in Community Health (SEARCH) study, a pair-matched, community randomized trial to estimate the effect of population-based HIV testing and streamlined ART on the 5-year cumulative HIV incidence (NCT01864603). The proposed methodology will be used in the primary analysis for the SEARCH trial. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: cluster randomized trials; pair-matching; population average treatment effect (PATE); sample average treatment effect (SATE); targeted maximum likelihood estimation (TMLE)

1. Introduction

In many studies, the goal is to estimate the impact of an exposure on the outcome of interest. Often, the target causal parameter is the population average treatment effect (PATE): the expected difference in the counterfactual outcomes if all members of some population were exposed and if all members of that population were unexposed. If there are no unmeasured confounders and there is sufficient variability in the exposure assignment (i.e., if the randomization and positivity assumptions hold), then we can identify the PATE as a function of the observed data distribution [1, 2]. The resulting statistical parameter can be estimated with a variety of algorithms, including matching and inverse weighting estimators (e.g., [1, 3, 4]), simple substitution estimators (e.g., [2, 5]), and double robust algorithms (e.g., [6–9]).

An alternative causal parameter is the sample average treatment effect (SATE) [10–15]. The sample effect is the average difference in the counterfactual outcomes for the actual study units. There are several potential advantages to selecting the SATE as the parameter of interest. First, the SATE is readily interpretable as the intervention effect for the sample at hand. Second, the SATE avoids assumptions about randomly sampling from and generalizing to some 'vaguely defined super-population of study units' [14]. In other words, the sample parameter remains relevant and interpretable if the units were systematically

[a] *Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, U.S.A.*
[b] *Division of Biostatistics, University of California, Berkeley, CA 94110-7358, U.S.A.*
[*] *Correspondence to: Laura Balzer, Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, U.S.A.*
[†] *E-mail: lbbalzer@hsph.harvard.edu*

selected for inclusion in the study, as is likely to be common in cluster randomized trials. Extensions of the study results to a broader or a different population can be addressed as a distinct research problem, approached with formal tools (e.g., [16–19]) and do not have to be assumed in the parameter specification. Finally, an estimator of the sample effect is often more precise than the same estimator of the population effect [10–13].

For a randomized trial, Neyman [10] first proposed estimating the SATE with the unadjusted estimator, which is the difference in the average outcomes among the treated units and the average outcomes among the control units. In this setting, the difference-in-means estimator will be unbiased for the SATE, conditional on the set of counterfactual outcomes for the study units. However, its variance remains unidentifiable as it relies on the correlation of the counterfactual outcomes [10–13]. Imbens [12] later generalized this work for an efficient estimator (i.e., a regular, asymptotically linear estimator, whose influence curve equals the efficient influence curve) in an observational setting. In particular, he showed that an efficient estimator for the population effect was unbiased for the sample effect, conditional on the baseline covariates and the counterfactual outcomes of the study units. He further expressed the variance of an efficient estimator of the SATE in terms of the variance of the same estimator of the PATE minus the variance of the unit-specific treatment effects across the population. This suggested that the standard variance estimator would be biased upwards unless there is no variability in the treatment effect.

Our contribution is to propose using targeted maximum likelihood estimation (TMLE) for estimation and inference of the sample effect in trials with and without pair-matching. TMLE is a general algorithm for constructing double robust, semiparametric, efficient, substitution estimators [8, 9]. Our results generalize the variance derivations of Imbens [12] to allow for misspecification of the outcome regression (i.e. the conditional mean outcome, given the exposure and covariates), estimation of the propensity score (i.e. the conditional probability of the receiving the exposure, given the covariates), and adaptive pair-matching [20]. Pair-matching is a popular design strategy in cluster randomized trials to protect study credibility and to increase power [20–25]. To the best of our knowledge, this is the first paper considering using an efficient estimator for the sample effect in a pair-matched trial.

We also contribute to the existing literature by formally defining each parameter, discussing interpretation, and examining identifiability within Pearl's structural causal model [26] as opposed to the Neyman–Rubin framework (e.g., [10–14]). Even though the SATE is formally not identified, we prove that the TMLE, presented here, is an asymptotically linear estimator of the SATE. Specifically, we show that the TMLE minus the sample effect behaves as an empirical mean of an influence curve depending on non-identifiable quantities and establish asymptotic normality with a non-identifiable limit variance. We propose a straightforward estimator of the upper bound of this variance, which nonetheless results in confidence intervals for the SATE that are smaller than those of the PATE. Simulations are used to evaluate the finite sample performance of our point estimator and proposed variance estimators. The simulations also serve to highlight the differences between the two causal parameters and the potential gains in power from selecting the sample effect as the target of inference and from pair-matching. Full R code for the simulations and estimators is provided in the Supporting Information. We motivate our discussion with the Sustainable East Africa Research in Community Health (SEARCH) trial for HIV prevention and treatment [27].

2. The causal model and causal parameters

SEARCH is an ongoing cluster randomized trial to evaluate the effect of a community-based strategy for HIV prevention and treatment in rural Uganda and Kenya (NCT01864603) [27]. In intervention communities, annual and targeted HIV testing is offered, and all individuals testing HIV+ are immediately eligible for antiretroviral therapy (ART) with streamlined delivery, including enhanced services for initiation, linkage, and retention in care. In control communities, all individuals testing HIV+ are offered ART according to the evolving in-country guidelines. The study hypothesis is that early HIV diagnosis combined with immediate and streamlined ART will reduce the 5-year cumulative HIV incidence. The primary outcome as well as other health, educational and economic outcomes will be measured among approximately 320,000 individuals, enrolled in the study. For the purposes of discussion, we focus on the community-level data. Thereby, our results are equally applicable to clustered and non-clustered data structures.

Consider the following data generating process for a randomized trial with two arms. First, the study units are selected. While some trials obtain a simple random sample from a well-defined target population, in other studies, there may not be a clear target population from which units were sampled and about which we wish to make inferences. In the SEARCH trial, for example, 32 communities were selected from Western Uganda (Mbarara region), Eastern Uganda (Tororo region), and the Southern Nyanza Province in Kenya by first performing ethnographic mapping on 54 candidate communities meeting the inclusion criteria (e.g., community size, health care infrastructure, and accessibility by a maintained transportation route), and then selecting the 16 pairs best matched on a range of characteristics (e.g., region, population density, occupational mix, and migration index) [20]. After selection of the study units, additional covariates are often measured. Additional covariates collected in the SEARCH trial included male circumcision coverage, measures of HIV prevalence, and measures of community-level HIV RNA viral load. Throughout the baseline covariates are denoted W.

Next, the intervention is randomized to the study units. Balanced allocation of the intervention can be guaranteed by randomly assigning the intervention to $n/2$ units and the control to remaining units or by randomizing within matched pairs. In the SEARCH trial, for example, the intervention was randomized within the 16 matched pairs. For ease of exposition, we present the causal model for the simple scenario, where the intervention is completely randomized, but our results are general. (Extensions to pair-matched trials are given in Section 5.) Let A be a binary variable, reflecting the assigned level of the intervention. For the SEARCH trial, A equals one if the community was assigned to the treatment (annual population-based testing and immediate and streamlined ART for all individuals testing HIV+) and equals zero if the community was assigned to the control (ART offered to HIV+ individuals according to in-country guidelines). At the end of follow-up, the outcome Y is measured. For the SEARCH trial, Y is the 5-year cumulative incidence of HIV. The observed data for a given study unit are then

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

Suppose we observe n independent, identically distributed (i.i.d.) copies of O with some distribution P_0. Throughout the subscript 0 will be used to denote the true distribution of the observed data. We note that for estimation and inference of the sample and conditional average treatment effects, we can weaken the i.i.d. assumption by conditioning on the vector of baseline covariates $(W_1, W_2, \ldots, W_n)$; for further details, see Balzer *et al.* [20].

The following structural causal model describes this data generating process [26, 28]. Each component of the observed data is assumed to be a deterministic function of its parents (variables that may influence its value) and unobservable background factors:

$$\begin{aligned} W &= f_W(U_W) \\ A &= \mathbb{I}(U_A < 0.5) \\ Y &= f_Y\left(W, A, U_Y\right) \end{aligned} \tag{1}$$

where the set of background factors $U = (U_W, U_A, U_Y)$ have some joint distribution P_U. By design, the random error determining the intervention assignment U_A is independent from the unmeasured factors contributing the baseline covariates U_W and the outcome U_Y:

$$U_A \perp\!\!\!\perp \left(U_W, U_Y\right).$$

Specifically, U_A is independently drawn from a Uniform(0,1). This causal model implies the statistical model for the set of possible distributions of the observed data O. In a randomized trial, the statistical model is semiparametric.

Through interventions on the structural causal model, we can generate the counterfactual outcome $Y(a)$, which is the outcome if possibly contrary-to-fact the unit was assigned $A = a$:

$$\begin{aligned} W &= f_W\left(U_W\right) \\ A &= a \\ Y(a) &= f_Y\left(W, a, U_Y\right). \end{aligned}$$

In this framework, the counterfactual outcomes $Y(a)$ are random variables. For the SEARCH trial, $Y(a)$ is the counterfactual cumulative incidence of HIV if possibly contrary-to-fact the community had been assigned treatment level $A = a$.

The distribution of the counterfactuals can then be used to define the causal parameter of interest. Often, the target of inference is the population average treatment effect:

$$PATE = \mathbb{E}\left[Y(1) - Y(0)\right].$$

This is the expected difference in the counterfactual outcomes for underlying target population from which the units were sampled. From the structural causal model, we see that the expectation is over the measured factors W and unmeasured factors U_Y, which determine the counterfactual outcomes for the population. In other words, the true value of the PATE does not depend on the sampled values of W or U_Y. For the SEARCH trial, the PATE would be the difference in the expected counterfactual cumulative incidence of HIV if possibly contrary-to-fact all communities in some hypothetical target population implemented the test-and-treat strategy, and the expected counterfactual cumulative incidence of HIV if possibly contrary-to-fact all communities in that hypothetical target population continued with the standard of care.

An alternative causal parameter is the sample average treatment effect, which was first proposed in Neyman [10]:

$$SATE = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i(1) - Y_i(0)\right].$$

This is simply the intervention effect for the n study units. The SATE is a data adaptive parameter; its value depends on the units included in the study. For recent work on estimation and inference of other data adaptive parameters, we refer the reader to [29, 30]. The SATE remains interpretable if there is no clear super-population from which the study units were selected. 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 $n = 32$ study communities.

In the SEARCH trial, targeting the sample effect has several advantages over targeting the population effect. First, there is no single real-world (as opposed to hypothetical) target population from which the study units were sampled or about which we wish to make inferences. While appropriate analytic approaches can reduce concerns over systematic sampling, the interpretation and policy relevance of the resulting PATE estimate would be unclear. In contrast, targeting the SATE allows us to rigorously estimate the intervention effect in a clearly defined, real-world population consisting of the roughly 320,000 persons resident in the 32 SEARCH communities. The resulting SATE estimate does not rely on any assumptions about the sampling mechanism, has a clear interpretation, and is generally more precise than an estimate of the PATE. As discussed in the succeeding text, estimators of the sample effect are at least as powerful as those of the population effect and expected to be more powerful when there is effect modification [11–13]. Clearly, however, it remains of significant policy interest to transport any effect found in the SEARCH trial to new populations and settings. However, alternative real-world target populations are likely to differ from the current setting in a number of ways that will likely impact the magnitude of the effect. As a result, neither the SATE nor the PATE will apply directly to these new settings. Thus, a desire for generalizability does not constitute an argument for favoring the PATE over the SATE. Instead, we argue that generalization (or transport) of the SEARCH effect to settings beyond the current sample is best addressed as a distinct research question, making full use of the modern toolbox available (e.g., [16–19]).

3. Identifiability

To identify the aforementioned causal effects, we must write them as some function of the observed data distribution P_0 [9, 12]. Under the randomization and positivity assumptions, we can identify the mean counterfactual outcome within strata of covariates [1, 2]:

$$\mathbb{E}\left[Y(a)|W\right] = \mathbb{E}\left[Y(a)|A = a, W\right] = \mathbb{E}_0\left[Y|A = a, W\right]$$

where the right-most expression is now in terms of the observed data distribution P_0. Briefly, the first equality holds under the randomization assumption, which states that the counterfactual outcome is independent of the exposure, given the measured covariates: $A \perp\!\!\!\perp Y(a)|W$. This is equivalent to the no unmeasured confounders assumption [1]. The positivity assumption states that the exposure level a occurs with a positive probability within all possible strata of covariates. Both assumptions hold by design in a randomized trial. As a well-known result, the PATE is identified as

$$\Psi^{\mathcal{P}}\left(P_0\right) = \mathbb{E}_0\left[\mathbb{E}_0(Y|A=1,W) - \mathbb{E}_0(Y|A=0,W)\right].$$

This statistical estimand is also called the G-computation identifiability result [2]. For the SEARCH trial, $\Psi^{\mathcal{P}}\left(P_0\right)$ would be the difference in expected cumulative HIV incidence, given the treatment and measured covariates, and the expected cumulative HIV incidence, given the control and measured covariates, averaged (standardized) with respect to the covariate distribution in the hypothetical target population. As with the causal parameter, there is one true value $\Psi^{\mathcal{P}}\left(P_0\right)$ for the population. In a randomized trial, conditioning on the covariates W is not needed for identifiability but will often provide efficiency gains during estimation (e.g., [31–37]).

In contrast, the SATE is not identifiable – in finite samples, we cannot strictly write the causal parameter as a function of the observed data distribution P_0. (Asymptotically, the SATE is identifiable, because the empirical mean converges to the expectation and thereby the sample effect converges to the population effect.) To elaborate, we can use the structural causal model (Eq. 1) to rewrite the sample effect as

$$\begin{aligned} SATE &= \frac{1}{n}\sum_{i=1}^{n}\left[Y_i(1) - Y_i(0)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} f_Y\left(W_i, 1, U_{Y_i}\right) - f_Y\left(W_i, 0, U_{Y_i}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[Y_i(1) - Y_i(0)\middle| W_i, U_{Y_i}\right]. \end{aligned}$$

The second equality is from the definition of counterfactuals as interventions on the causal model. The final equality is the conditional average treatment effect (CATE), given the measured baseline covariates as well as the unmeasured factors. The conditional effect was first proposed in Abadie and Imbens [38] and is the average difference in the expected counterfactual outcomes, treating the measured covariates of the study units as fixed: CATE$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[Y_i(1) - Y_i(0)\middle| W_i\right]$. This representation of the SATE suggests that if we had access to all pre-intervention covariates impacting the outcome (i.e., $\{W, U_Y\}$), then we could apply the results for estimation and inference for the conditional parameter, as detailed in Balzer *et al.* [20]. In reality, we only measure a subset of these covariates (i.e., W), and only this subset is available for estimation and inference. Therefore, the SATE is formally not identifiable in finite samples. Nonetheless, as detailed later, a TMLE developed for the population effect will be consistent and asymptotically linear for the sample effect, and the corresponding variance estimator will be asymptotically conservative.

4. Estimation and inference

There are many well-established algorithms for estimation of the population parameter $\Psi^{\mathcal{P}}\left(P_0\right)$. Examples include inverse probability of treatment weighting, simple substitution estimators, augmented inverse probability of treatment weighting, and TMLE (e.g., [1–9]). In a randomized trial, the unadjusted difference in the average outcomes among the treated units and the average outcome among the control units provides a simple and unbiased estimate of the PATE. Adjusting for measured covariates, however, will generally increase efficiency and study power (e.g., [4, 31–37].) For example, we can obtain a more precise estimator of the PATE by (i) regressing the outcome Y on the exposure A and covariates W; (ii) using the estimated coefficients to obtain the predicted outcomes for all units under the exposure and control; and (iii) then taking the average difference in the predicted outcomes. For a large class of general linear models, there is no risk of bias if the 'working' model for the outcome regression is misspecified [36]. This algorithm is called parametric G-computation [2] in observational studies and also called analysis of covariance [32] in the special case of a continuous outcome and a linear model without interactions. Alternatively, we can obtain a more precise estimator of $\Psi^{\mathcal{P}}\left(P_0\right)$ by estimating known exposure mechanism to capture chance imbalances in the covariate distribution between treatment groups (e.g., [4, 7, 35]).

In the SEARCH trial, for example, the true conditional probability of being assigned to the test-and-treat intervention is $P_0(A = 1|W) = 0.5$. However, with only $n = 32$ communities, there is likely to be variation in the baseline covariates across the treatment arms.

We focus our discussion on TMLE, which incorporates estimation of both the outcome regression (the conditional mean outcome given the exposure and covariates) and the propensity score (the conditional probability of receiving the exposure given the covariates [1]). In general, TMLE is a double robust estimator; it will be consistent if either outcome regression or the propensity score is consistently estimated. If both functions are consistently estimated at a fast enough rate and there is sufficient variability in the propensity score, the estimator is also asymptotically efficient in that it attains the lowest possible variance among a large class of regular, asymptotically linear estimators. TMLE is also a substitution (plug-in) estimator, which provides stability in the context of sparsity [39,40]. Finally, TMLE makes use of state-of-the-art machine learning and therefore avoids the parametric assumptions commonly made in other algorithms. In other words, TMLE does not place any unwarranted assumptions on the structure of the data and respects the semiparametric statistical model.

4.1. Targeted maximum likelihood estimation for the population effect

For the population parameter $\Psi^{\mathcal{P}}\left(P_0\right)$, a TMLE can be implemented as follows.

- *Step 1. Initial estimation:* First, we obtain an initial estimator of the outcome regression $\mathbb{E}_0(Y|A, W)$. For example, the outcome Y can be regressed on the exposure A and covariates W according to a parametric 'working' model [36]. Alternatively, we could use an *a priori* specified data adaptive procedure, such as SuperLearner [41].
- *Step 2. Targeting:* Second, we update the initial estimator of the outcome regression $\mathbb{E}_n(Y|A, W)$ by incorporating information on the exposure-covariate relation (i.e., the propensity score). Informally, this 'targeting' step helps to remove some of the residual imbalance in the baseline covariate distributions across treatment groups. More formally, this targeting step serves to obtain the optimal bias-variance trade-off for $\Psi^{\mathcal{P}}\left(P_0\right)$ and to solve the efficient score equation [42]. The reader is referred to van der Laan and Rose [9] for further details. This targeting step is implemented as follows.
  - We calculate the 'clever covariate' [9] based on the known or estimated propensity score $P_n(A = 1|W)$:

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

    (To estimate the propensity score, we could run logistic regression of the exposure A on the covariates W or use more data adaptive methods.)
  - For a continuous and unbounded outcome, we run linear regression of the outcome Yon the covariate $H_n(A, W)$ with the initial estimator as offset (i.e., we suppress the intercept and set the coefficient on the initial estimator equal to 1). We plug in the estimated coefficient ϵ_n to yield the targeted update: $\mathbb{E}^*_n(Y|AW) = E_n(Y|A, W) + \epsilon_n H_n(A, W)$.
  - For a binary or a bounded continuous outcome (e.g., a proportion) [39], we run logistic regression of the outcome Y on the covariate $H_n(A, W)$ with the $logit(\cdot) = log[\cdot/(1 - \cdot)]$ of the initial estimator as offset. We plug in the estimated coefficient ϵ_n to yield the targeted update: $\mathbb{E}^*_n(Y|A, W) = logit^{-1}\left\{logit\left[\mathbb{E}_n(Y|A, W)\right] + \epsilon_n H_n(A, W)\right\}$.
- *Step 3. Parameter estimation:* Finally, we obtain a point estimate by substituting the targeted estimates into the parameter mapping:

$$\Psi_n\left(P_n\right) = \frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{E}^*_n(Y_i|A_i = 1, W_i) - \mathbb{E}^*_n(Y_i|A_i = 0, W_i)\right]$$

  where P_n denotes the empirical distribution, placing mass $1/n$ on each observation O_i. The sample mean is the nonparametric maximum likelihood estimator of the marginal distribution of the baseline covariates $P_0(W)$.

We note if the propensity score is not estimated and the working regression model used for initial estimation of $\mathbb{E}_0(Y|A, W)$ contains an intercept and a main term for the exposure, then this targeting step will not yield an update and can be skipped [35, 36].

Under standard regularity conditions, this TMLE is a consistent and asymptotically linear estimator of the population parameter [8, 9]:

$$\Psi_n\left(P_n\right) - \Psi^{\mathcal{P}}\left(P_0\right) = \frac{1}{n}\sum_{i=1}^{n} D^{\mathcal{P}}\left(O_i\right) + o_P\left(1/\sqrt{n}\right).$$

In words, the estimator minus the truth can be written as an empirical mean of an influence curve $D^{\mathcal{P}}(O)$ and a second-order term going to 0 in probability. The influence curve is given by

$$\begin{aligned} D^{\mathcal{P}}(O) &= D_Y(O) + D_W(O) \\ D_Y(O) &= \left(\frac{\mathbb{I}(A=1)}{P_0(A=1|W)} - \frac{\mathbb{I}(A=0)}{P_0(A=0|W)}\right)\left(Y - \mathbb{E}^*_{n,lim}(Y|A,W)\right) \\ D_W(O) &= \mathbb{E}^*_{n,lim}(Y|A=1,W) - \mathbb{E}^*_{n,lim}(Y|A=0,W) - \Psi^{\mathcal{P}}\left(P_0\right) \end{aligned}$$

where $\mathbb{E}^*_{n,lim}(Y|A, W)$ denotes the limit of the TMLE $\mathbb{E}^*_n(Y|A, W)$ and we are assuming the propensity score is known or consistently estimated, as will always be true when the treatment A is randomized. The first term of the influence curve D_Y is the weighted residuals (i.e., the weighted deviations between the observed outcome and the limit of the predicted outcome). The second term D_W is deviation between the limit of the estimated strata-specific association and the marginal association.

The standardized estimator is asymptotically normal with variance given by the variance of its influence curve $D^{\mathcal{P}}(O)$, divided by sample size n [8, 9]. Under consistent estimation of the outcome regression (i.e., when $\mathbb{E}^*_{n,lim}(Y|A, W) = \mathbb{E}_0(Y|A, W)$), the TMLE will be asymptotically efficient and achieve the lowest possible variance among a large class of estimators of the population effect. In other words, its influence curve will equal the efficient influence curve, and the TMLE will achieve the efficiency bound of Hahn [42]. Thereby, improved estimation of the outcome regression leads to more precise estimators of the population effect. In finite samples, the variance of the TMLE is well approximated by the sample variance of the estimated influence curve scaled by sample size:

$$\sigma_n^{2,\mathcal{P}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[D_n^{\mathcal{P}}\left(O_i\right)\right]^2}{n} \tag{2}$$

where

$$\begin{aligned} D_n^{\mathcal{P}}(O) = &\left(\frac{\mathbb{I}(A=1)}{P_n(A=1|W)} - \frac{\mathbb{I}(A=0)}{P_n(A=0|W)}\right)\left(Y - \mathbb{E}^*_n(Y|A,W)\right) \\ &+ \mathbb{E}^*_n(Y|A=1,W) - \mathbb{E}^*_n(Y|A=0,W) - \Psi_n\left(P_n\right). \end{aligned}$$

The algorithm is available in the `tmle` [43] and `ltmle` [44] packages in `R` [45]. Full `R` code is also given in Appendix D of the Supporting Information.

4.2. *Targeted maximum likelihood estimation for the sample effect*

For a randomized trial, Neyman [10] proposed estimating the SATE with the unadjusted estimator:

$$\Psi_{n,unadj}\left(P_n\right) = \frac{\sum_{i=1}^{n}\mathbb{I}(A_i=1)Y_i}{\sum_{i=1}^{n}\mathbb{I}(A_i=1)} - \frac{\sum_{i=1}^{n}\mathbb{I}(A_i=0)Y_i}{\sum_{i=1}^{n}\mathbb{I}(A_i=0)}.$$

Conditional on the vector of counterfactual outcomes $\mathbf{Y}(\mathbf{a}) = \{Y_i(a) : i = 1, \ldots, n,\ a = 0, 1\}$, the difference-in-means estimator is unbiased but inefficient. To the best of our knowledge, Imbens [12] was the first to discuss an efficient estimator (i.e., a regular, asymptotically linear estimator, whose influence curve equals the efficient influence curve) of the sample effect. He proved that an efficient estimator for the PATE was unbiased for the SATE, given the vector of baseline covariates $\mathbf{W} = (W_1, \dots, W_n)$ and the set of counterfactual outcomes $\mathbf{Y(a)} = \{Y_i(a) : i = 1, \dots, n, \; a = 0, 1\}$. We now extend these results to TMLE. Specifically, we allow the estimator of outcome regression $\mathbb{E}_0(Y|A,W)$ to converge to a possibly misspecified limit, incorporate estimation of the known propensity score, and suggest an alternate method for variance estimation. In Section 5, we further extend these results to a pair-matched trial.

The TMLE for the population parameter $\Psi^{\mathcal{P}}\left(P_0\right)$, presented in Section 4.1, also serves as an estimator of the SATE. The implementation is identical. Furthermore, under typical regularity conditions, the TMLE minus the sample effect behaves as an empirical mean of an influence curve depending on non-identifiable quantities, and a second-order term going to zero in probability:

$$\Psi_n\left(P_n\right) - SATE = \frac{1}{n}\sum_{i=1}^{n} D^{\mathcal{S}}\left(U_i, O_i\right) + o_P\left(1/\sqrt{n}\right)$$

where

$$\begin{aligned} D^{\mathcal{S}}\left(U_i, O_i\right) &= D^{\mathcal{C}}\left(O_i\right) - D^{\mathcal{F}}\left(U_i, O_i\right) \\ D^{\mathcal{C}}\left(O_i\right) &= D_Y\left(O_i\right) - \mathbb{E}_0\left[D_Y\left(O_i\right)|\mathbf{W}\right] \end{aligned} \tag{3}$$

$$D^{\mathcal{F}}\left(U_i, O_i\right) = Y_i(1) - Y_i(0) - \left[\mathbb{E}_0\left(Y_i|A_i = 1, W_i\right) - \mathbb{E}_0\left(Y_i|A_i = 0, W_i\right)\right]. \tag{4}$$

(Proof in Appendix A of the Supporting Information.) The first component $D^{\mathcal{C}}$ is the influence curve for the TMLE of the conditional parameter $\Psi^{\mathcal{C}}\left(P_0\right) = 1/n\sum_{i=1}^{n}[\mathbb{E}_0(Y_i|A_i = 1, W_i) - \mathbb{E}_0(Y_i|A_i = 0, W_i)]$, which corresponds to the conditional average treatment effect (CATE) under the necessary identifiability assumptions [20]. This term depends on the true outcome regression $\mathbb{E}_0(Y|A,W)$. Specifically, the conditional expectation of the D_Y component, given the baseline covariates, equals the deviation between the true conditional means and the limits of the estimated conditional means:

$$\begin{aligned} \mathbb{E}_0\left[D_Y(O)|\mathbf{W}\right] = &\left[\mathbb{E}_0(Y|A=1,W) - \mathbb{E}_0(Y|A=0,W)\right] \\ &- \left[\mathbb{E}^*_{n,lim}(Y|A=1,W) - \mathbb{E}^*_{n,lim}(Y|A=0,W)\right]. \end{aligned}$$

Under consistent estimation of the outcome regression (i.e., when $\mathbb{E}^*_{n,lim}(Y|A,W) = \mathbb{E}_0(Y|A,W)$), this term is zero. The second component $D^{\mathcal{F}}$ is a function of the unobserved factors $U = (U_W, U_A, U_Y)$ and the observed data $O = (W, A, Y)$. This non-identifiable term captures the deviations between the unit-specific treatment effect and expected effect within covariate strata:

$$\begin{aligned} D^{\mathcal{F}}\left(U_i, O_i\right) &= Y_i(1) - Y_i(0) - \left[\mathbb{E}_0\left(Y_i|A_i = 1, W_i\right) - \mathbb{E}_0\left(Y_i|A_i = 0, W_i\right)\right] \\ &= Y_i(1) - Y_i(0) - \left[\mathbb{E}\left(Y_i(1)|W_i\right) - E\left(Y_i(0)|W_i\right)\right] \\ &= Y_i(1) - Y_i(0) - \mathbb{E}\left[Y_i(1) - Y_i(0)\,|W_i\right]. \end{aligned}$$

In the last line, the expectation is over the unmeasured factors U_Y that determine the counterfactual outcomes. This term will be zero if there is no variability in the treatment effect across units with the same values of the measured covariates. We also note that there is no contribution to the influence curve $D^{\mathcal{S}}$ from estimation of the covariate distribution, which is considered fixed. In other words, there is no D_W component to the influence curve.

As a result, the standardized estimator of the SATE is consistent and asymptotically normal with mean zero and variance given by the limit of

$$\begin{aligned} Var[D^{\mathcal{S}}(U,O)] &= Var\left[D^{\mathcal{C}}(O)\right] + Var\left[D^{\mathcal{F}}(U,O)\right] - 2Cov\left[D^{\mathcal{C}}(O), D^{\mathcal{F}}(U,O)\right] \\ &= Var\left[D^{\mathcal{C}}(O)\right] - Var\left[D^{\mathcal{F}}(U,O)\right]. \end{aligned}$$

(Proof in Appendix A.1 of the Supporting Information.) Because the variance of the non-identifiable $D^{\mathcal{F}}$ component must be greater than or equal to zero, the asymptotic variance of the TMLE as an estimator of the sample effect will always be less than or equal to the asymptotic variance of the same estimator of the conditional effect. They will only have the same precision when there is no variability in the unit-level treatment effect within strata of measured covariates (i.e., when $Var[D^{\mathcal{F}}(U,O)] = 0$). In many settings, however, there will be heterogeneity in the effect, and the TMLE for the SATE will be more precise. Even if the treatment effect is constant within covariate strata, the TMLE for the sample effect (or the conditional effect) will always be at least as precise as the same TMLE for the population effect. They will only have the same efficiency bound when (i) the outcome regression is consistently estimated; (ii) there is no variability in the treatment effect *across* strata of measured covariates (i.e., when $Var[D_W(O)] = 0$); and (iii) there is no variability in the treatment effect *within* strata of measured covariates. In many settings, there will be effect modification, and focusing on estimation of the SATE will yield the most precision and power.

We can conservatively approximate the influence curve for the TMLE of the sample effect as

$$D_n^{\mathcal{S}}\left(O_i\right) = D_{Y,n}\left(O_i\right) = \left(\frac{\mathbb{I}(A_i = 1)}{P_n(A_i = 1|W_i)} - \frac{\mathbb{I}(A_i = 0)}{P_n(A_i = 0|W_i)}\right)\left(Y_i - \mathbb{E}_n^*(Y_i|A_i, W_i)\right). \tag{5}$$

(Further details in Appendix A.1–A.2 of the Supporting Information.) Thereby, we obtain an asymptotically conservative variance estimator with the sample variance of the weighted residuals scaled by sample size n:

$$\sigma_n^{2,\mathcal{S}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[D_n^{\mathcal{S}}\left(O_i\right)\right]^2}{n}. \tag{6}$$

As for the PATE, adjusting for predictive baseline covariates can substantially improve power for the SATE by reducing variability in the estimator. Unlike the PATE, however, adjusting for predictive baseline covariates can provide an additional power gain for the SATE by resulting in a less conservative variance estimator. Furthermore, this variance estimator is easy to implement as the relevant pieces are known or already estimated. As a result, this may provide an attractive alternative to the matching estimator of the variance, proposed by Abadie and Imbens [38] and discussed in Imbens [12]. We note that the bootstrap is inappropriate as the SATE changes with each sample. Fisher's permutation distribution is also not appropriate, because it is testing the strong null hypothesis of no treatment effect for any unit ($Y_i(1) = Y_i(0)$, $\forall i$) [46], whereas our interest is in the weak null hypothesis of no average treatment effect.

5. Extensions to pair-matched trials

We recall that the SEARCH trial is a pair-matched study. Briefly, $N = 54$ candidate communities, satisfying the study's inclusion criteria, were identified. Of these, the best $n/2 = 16$ matched pairs were chosen according to similarity on the baseline covariates of the candidate units. This 'adaptive pair-matching' scheme is detailed in Balzer *et al.* [20] and also called 'nonbipartite matching' and 'optimal multivariate matching' in other contexts [22, 47, 48]. This study design creates a dependence in the data. Specifically, the construction of the matched pairs is a function of the covariates of all candidate sites. As a result, the observed data cannot be treated as n i.i.d. observations nor as $n/2$ i.i.d. paired observations, as current practice sometimes assumes (e.g., [21, 24, 49, 50]). However, once the baseline covariates of the study units are considered to be fixed, we recover $n/2$ conditionally independent units:

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

where the index $j = 1, \ldots, n/2$ denotes the partitioning of the candidate study communities $\{1, \ldots, N\}$ into matched pairs according to their baseline covariates $(W_1, \ldots, W_N)$.

Previously, Imai [13] generalized Neyman's analysis of the unadjusted estimator for the sample effect in a pair-matched trial. The unadjusted estimator, as the average of the pairwise differences in outcomes, is unbiased but inefficient. For an adaptive pair-matched trial, van der Laan *et al.* [25] detailed the use TMLE for the population effect, and Balzer *et al.* [20] for the conditional effect. To the best of our knowledge, this is the first paper to consider using a locally efficient estimator for the sample effect in a pair-matched trial.

5.1. Targeted maximum likelihood estimation for the sample effect in a pair-matched trial

The TMLE for the population effect, presented in Section 4.1, also estimates the sample effect in a pair-matched trial. As before, the TMLE minus the SATE can be written as an empirical mean of an influence curve depending on non-identifiable quantities, and a second-order term going to zero in probability:

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

where

$$\begin{aligned}
\bar{D}^{\mathcal{S}}\left(\bar{U}_j, \bar{O}_j\right) &= \bar{D}^{\mathcal{C}}\left(\bar{O}_j\right) - \bar{D}^{\mathcal{F}}\left(\bar{U}_j, \bar{O}_j\right)\\
\bar{D}^{\mathcal{C}}\left(\bar{O}_j\right) &= \frac{1}{2}\left[D^{\mathcal{C}}\left(O_{j1}\right) + D^{\mathcal{C}}\left(O_{j2}\right)\right]\\
\bar{D}^{\mathcal{F}}\left(\bar{U}_j, \bar{O}_j\right) &= \frac{1}{2}\left[D^{\mathcal{F}}\left(U_{j1}, O_{j1}\right) + D^{\mathcal{F}}\left(U_{j2}, O_{j2}\right)\right].
\end{aligned}$$

(Proof in Appendix B of the Supporting Information.) The first component $\bar{D}^{\mathcal{C}}(\bar{O})$ is the influence curve for the TMLE of the conditional parameter $\Psi^{\mathcal{C}}\left(P_0\right) = 1/n\sum_{i=1}^{n} \mathbb{E}_0(Y_i|A_i = 1, W_i) - \mathbb{E}_0(Y_i|A_i = 0, W_i)$ in a trial with pair-matching [20]. In words, $\bar{D}^{\mathcal{C}}\left(\bar{O}_j\right)$ is the average of the pairwise $D^{\mathcal{C}}\left(O_i\right)$ components, as defined in Eq. (3). The second component $\bar{D}^{\mathcal{F}}\left(\bar{U}, \bar{O}\right)$ is a non-identifiable function of the pair's unobserved factors $\bar{U} = (U_{j1}, U_{j2})$ and observed factors $\bar{O}_j = (O_{j1}, O_{j2})$. Specifically, $\bar{D}^{\mathcal{F}}\left(\bar{U}_j, \bar{O}_j\right)$ is the average of the pairwise $D^{\mathcal{F}}\left(U_i, O_i\right)$ components, as defined in Eq. (4). As before, there is no contribution from estimation of the covariate distribution $P_0(W)$, which is considered fixed.

As a consequence, the standardized estimator of the SATE in a pair-matched trial is consistent and asymptotically normal with mean zero and variance given by the limit of

$$Var\left[\bar{D}^{\mathcal{S}}\left(\bar{U}_j, \bar{O}_j\right)\right] = Var\left[\bar{D}^{\mathcal{C}}\left(\bar{O}_j\right)\right] - Var\left[\bar{D}^{\mathcal{F}}\left(\bar{U}_j, \bar{O}_j\right)\right]$$

(Proof in Appendix B.1 of the Supporting Information). As before, the variance of the non-identifiable $\bar{D}^{\mathcal{F}}$ component must be greater than or equal to zero. Therefore, in a pair-matched trial, the asymptotic variance of the TMLE as an estimator of the sample effect will always be less than or equal to the asymptotic variance of the same estimator of the conditional effect. Furthermore, by treating the covariate distribution as fixed, the TMLE for the sample (or conditional) effect will always be as or more precise than the TMLE of the population effect in a pair-matched trial. We also briefly note that there is often an additional efficiency gain due to pair-matching (Appendix C of the Supporting Information). The SATE will be estimated with more precision in a pair-matched trial when the deviations between the true and estimated outcome regressions are positively correlated within matched pairs and/or when the deviations between the treatment effect for a unit and the treatment effect within covariate strata are positively correlated within matched pairs.

We can conservatively approximate the influence curve for the TMLE of the SATE in a pair-matched trial as

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

where $D_n^{\mathcal{S}}\left(O_i\right)$ is defined in Eq. (5) (Appendix B.1 of the Supporting Information). Thereby, we obtain an asymptotically conservative variance estimator with the sample variance of the estimated paired influence curve, divided by sample size $n/2$:

$$\bar{\sigma}_n^{2,\mathcal{S}} = \frac{\frac{1}{n/2}\sum_{j=1}^{n/2}\left[\bar{D}_n^{\mathcal{S}}\left(\bar{O}_j\right)\right]^2}{n/2}. \tag{7}$$

If we order 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 ($A_{j2} = 0$) and treat the exposure mechanism as known $P_0(A) = 0.5$, it follows that

$$\bar{D}_n^{\mathcal{S}}\left(\bar{O}_j\right) = \left(Y_{j1} - \mathbb{E}_n^*\left(Y_{j1}|A_{j1}, W_{j1}\right)\right) - \left(Y_{j2} - \mathbb{E}_n^*\left(Y_{j2}|A_{j2}, W_{j2}\right)\right). \tag{8}$$

In this case, we can represent the variance estimator as the sample variance of the difference in residuals within matched pairs, divided by $n/2$. This variance estimator will be consistent if there is no heterogeneity in the treatment effect within strata of measured covariates (i.e., if the variance of the $\bar{D}^{\mathcal{F}}$ component is zero) *and* if the outcome regression $\mathbb{E}_0(Y|A,W)$ is consistently estimated. Under the same conditions, the TMLE will be efficient (i.e., achieve the lowest possible variance among a large class of regular, asymptotically linear estimators). Otherwise, the TMLE will not be efficient, and the variance estimator will be conservative. As before, adjusting for predictive baseline covariates can substantially improve power in two ways: (i) by reducing variability in the estimator and (ii) by resulting in a less conservative variance estimator.

6. Simulation study

We present the following simulation study to (i) further illustrate the differences between the causal parameters; (ii) demonstrate implementation of the TMLE; and (iii) understand the impact of the parameter specification on the estimator's precision and attained power. We focus on a randomized trial to illustrate the potential gains in efficiency with pair-matching during the design and with adjustment during the analysis. All simulations were carried out in R v3.2.2 [45]. Full R code is available in Appendix D of the Supporting Information.

6.1. Data generating process and estimators

Consider the following data generating process for unit $i=\{1,\dots,n\}$. First, we generated the background error $U_{Y,i}$ by drawing from a standard normal distribution. Then, we generated five baseline covariates from a multivariate normal with means 0 and standard deviation 1. The correlation between the first two covariates $(W1_i, W2_i)$ was 0, and the correlation between the last three $(W3_i, W4_i, W5_i)$ was 0.65. The exposure A_i was randomized such that the treatment allocation was balanced overall. Recall A_i is a binary indicator, equaling 1 if the unit is randomized to the intervention and 0 if the unit is randomized to the control. For a trial without matching, the intervention was randomly assigned to $n/2$ units and the control to the remaining units. For a trial with matching, we applied the nonbipartite matching algorithm `nbpMatch` [51] to pair units on $\{W1, W4, W5\}$. The outcome Y_i was generated as

$$Y_i = logit^{-1}\left[A_i + 0.75W1_i + 0.75W2_i + 1.25W3_i + U_{Y,i} + 0.75A_iW1_i - 0.5A_iW2_i - A_iU_{Y,i}\right]/5.$$

We also generated the counterfactual outcomes $Y_i(a)$ by intervening to set $A_i = a$. For sample sizes of $n = \{30, 50\}$, this data generating process was repeated 5000 times. The true value of the SATE was calculated as the average difference in the counterfactual outcomes for each sample, and the true value of the PATE was calculated by averaging the difference in the counterfactual outcomes over a population of 500,000 units. In this population, the correlations between the observed outcome Y and the baseline covariates were weak to moderate: 0.5 for $W1$, 0.2 for $W2$, 0.6 for $W3$, 0.4 for $W4$ and 0.4 for $W5$.

We compared the performance of the unadjusted estimator with the TMLE with two methods for initial estimation of the outcome regression. Specifically, we estimated $\mathbb{E}_0(Y|A,W)$ with logistic regression, including as main terms the exposure A, the covariate $W1$ and an interaction A^*W1. We also estimated $\mathbb{E}_0(Y|A,W)$ with SuperLearner, an optimal machine-learning approach [41]. In particular, we used cross-validation to create the best convex combination of algorithm-specific estimates from a pre-specified library, which consisted of all possible logistic regressions with terms for the exposure A, a single covariate and their interaction. The unadjusted estimator can be considered as a special case of the TMLE, where $\mathbb{E}_n(Y|A,W) = \mathbb{E}_n(Y|A)$. Inference was based on the estimated influence curve and the Student's t-distribution. We constructed Wald-type 95% confidence intervals and tested the null hypothesis of no average effect.

6.2. Simulation results

Table I gives a summary of the parameter values across the 5000 simulated trials. Recall the true value of the SATE depends on the units included in the study, whereas there is one true value of the PATE for the population. The sample effect ranged from 0.17% to 5.94% with a mean of 2.97%. The population effect was constant at 2.98%. As expected, the variability in the SATE decreased with increasing sample size.

Table I. Summary of the causal parameters (in %) over 5000 simulations of size $n = \{30, 50\}$.

	SATE				PATE			
	min	ave	max	var	min	ave	max	var
$n = 30$	0.17	2.97	5.94	6.5E-3	2.98	2.98	2.98	0
$n = 50$	0.18	2.96	5.14	4.2E-3	2.98	2.98	2.98	0

SATE, sample average treatment effect; PATE, population average treatment effect; var, variance of the causal parameter over 5000 simulations.

Table II. Summary of estimator performance over 5000 simulations.

Target and design	Estimator	Bias	σ	MSE	rMSE	Power	Coverage
			Sample size $n = 30$				
PATE and not matched	Unadj	2.3E-4	2.2E-2	4.8E-4	1.00	0.27	0.95
	TMLE	6.8E-4	1.9E-2	3.6E-4	0.75	0.36	0.94
	TMLE+SL	2.9E-4	1.6E-2	2.6E-4	0.55	0.48	0.93
SATE and not matched	Unadj	3.1E-4	2.0E-2	4.2E-4	0.88	0.27	0.96
	TMLE	7.5E-4	1.7E-2	3.0E-4	0.63	0.39	0.95
	TMLE+SL	3.7E-4	1.4E-2	2.0E-4	0.42	0.52	0.95
SATE and matched	Unadj	5.4E-5	1.5E-2	2.2E-4	0.46	0.37	0.98
	TMLE	3.7E-4	1.4E-2	2.1E-4	0.43	0.44	0.97
	TMLE+SL	1.3E-4	1.1E-2	1.3E-4	0.27	0.58	0.97
			Sample Size $n = 50$				
PATE and not matched	Unadj	−1.3E-4	1.7E-2	3.0E-4	1.00	0.41	0.94
	TMLE	1.1E-4	1.5E-2	2.2E-4	0.75	0.53	0.94
	TMLE+SL	−3.1E-6	1.2E-2	1.6E-4	0.53	0.68	0.94
SATE and not matched	Unadj	4.8E-5	1.6E-2	2.5E-4	0.86	0.41	0.96
	TMLE	2.9E-4	1.3E-2	1.8E-4	0.60	0.55	0.96
	TMLE+SL	1.8E-4	1.1E-2	1.1E-4	0.38	0.70	0.97
SATE and matched	Unadj	−1.8E-4	1.1E-2	1.1E-4	0.38	0.59	0.98
	TMLE	−1.6E-4	1.0E-2	1.1E-4	0.36	0.66	0.97
	TMLE+SL	−5.7E-5	8.2E-3	6.7E-5	0.23	0.81	0.98

The rows denote target parameter, the study design, and the estimator: unadjusted, TMLE with logistic regression, and TMLE with SuperLearner (‘TMLE+SL’). The columns denote estimator performance with σ as the standard deviation of the estimator for its target, and rMSE as the MSE of an estimator divided by the MSE of the unadjusted estimator of the PATE in a trial without matching.

PATE, population average treatment effect; SATE, sample average treatment effect; MSE, mean squared error; rMSE, relative MSE; TMLE, targeted maximum likelihood estimation.

Table II illustrates the performance of the estimators. Specifically, we give the bias as the average deviation between the point estimate and (sample-specific) true value, the standard deviation σ as the square root of the variance of an estimator for its target, and the mean squared error (MSE). We also show the relative MSE as the MSE of a given estimator divided by the MSE of the unadjusted estimator of the population effect in trial without matching. The attained power, which is the proportion of times the false null hypothesis was rejected, and the 95% confidence interval coverage are also included.

As expected, all estimators were unbiased. In randomized trials, there is no risk of bias because of misspecification of the regression model for $\mathbb{E}_0(Y|A, W)$ (e.g., [34–36]). Also as expected, the precision of the estimators improved with increasing sample size and with adjustment (e.g., [31–36]). Consider, for example, estimation of the population effect in a trial with $n = 30$ units and without matching. The standard error was 2.2^*10^{-2} for the unadjusted estimator and 1.9^*10^{-2} after adjusting for a single covariate. Incorporating data adaptive estimation of the conditional mean $\mathbb{E}_0(Y|A, W)$ through SuperLearner further reduced the standard error to 1.6^*10^{-2}. Also as expected, precision increased with pair-matching [20, 23, 25] (Appendix C of the Supporting Information). For the SATE, the standard error of the unadjusted estimator in the trial without matching was 1.38 times higher with $n = 30$ units and 1.49 times higher with $n = 50$ units than its pair-matched counterpart.

For all estimation algorithms and sample sizes, the impact of the target parameter specification on precision and power was substantial. As predicted by theory, the highest variance was seen with the unadjusted estimator of the PATE. With $n = 50$ units, the MSE of this estimator for the PATE was 2.62 times that of the TMLE with SuperLearner for the SATE in a trial without matching and 4.42 times that of the TMLE with SuperLearner for the SATE in a trial with matching. In the finite sample simulations, the impact of having an asymptotically conservative variance estimator on inference for sample effect was notable. In most settings, the standard deviation of an estimator of the SATE was over-estimated, and the confidence interval coverage was greater than or equal to the nominal rate of 95%. Despite the conservative variance estimator, the TMLE for the sample effect achieved higher power than the same TMLE for the population effect. With $n = 30$ units, the attained power for the TMLE with SuperLearner was 48% for the population effect, 52% for the sample effect without matching, and 58% for the sample effect after pair-matching. With $n = 50$ units, the attained power for the TMLE with SuperLearner was 68% for the population effect, 70% for the sample effect without matching, and 81% for the sample effect after pair-matching. Notably, the power was the same for the unadjusted estimator of the two parameters in the trials without matching. The power of the unadjusted estimator did not vary, because the estimated $D_W(O)$ component of influence curve and thereby its variance were zero:

$$\mathbb{E}_n(Y|A=1) - \mathbb{E}_n(Y|A=0) - \Psi_{n,unadj}\left(P_n\right) = 0$$

where $\mathbb{E}_n(Y|A)$ denotes the treatment-specific mean. Thus, using the unadjusted estimator sacrificed any potential gains in power by specifying the SATE as the target of inference. In contrast, the TMLE using SuperLearner was able to obtain a better fit of the outcome regression $\mathbb{E}_0(Y|A, W)$ and a less conservative variance estimator. As a result, this TMLE was able to achieve the most power.

7. Discussion

This work was motivated by the SEARCH trial for HIV prevention and treatment [27]. The SEARCH trial will capture the effect of a community-based strategy for immediate and streamlined ART on ≈ 320,000 people in rural Uganda and Kenya. For the following reasons, the SATE was chosen as the target of inference for the primary analysis. The candidate communities were systematically selected to satisfy the study's inclusion criteria and then a matching algorithm applied to select the best 16 matched pairs [20]. Therefore, the observed data did not arise from taking a simple random sample from some hypothetical target population of matched pairs of communities. In this setting, the SATE, in contrast to the PATE, remains a readily interpretable quantity that can be rigorously estimated without further assumptions on the sampling mechanism. While generalizability of the study findings and their transport to new settings remains of substantial policy interest, neither the SATE nor the PATE directly addresses this goal; these new settings are likely to differ in important ways from both the current sample and any hypothetical target population from which it was drawn. Instead, we advocate approaching generalizability and transportability as distinct research questions, requiring their own identification results and corresponding optimal estimators [14, 16–19]. Finally, the sample effect will be estimated with at least as much precision and power as the conditional or population effects.

To our knowledge, this is the first paper to propose using TMLE for estimation and inference of the SATE in trials with and without pair-matching. Despite the lack of identifiability of the SATE in finite samples, we proved that the TMLE was a consistent and asymptotically normal estimator of the SATE. If there is heterogeneity in the intervention effect within strata of measured covariates or across strata of measured covariates, the sample effect will be estimated with more precision than the population effect. We also provided asymptotically conservative variance estimators, which are intuitive and straightforward to implement. Furthermore, we showed that a trial targeting the sample effect and implementing adaptive pair-matching will often be more efficient than a trial targeting the sample effect and not implementing pair-matching.

Finite sample simulations highlighted the differences between the causal parameters and the impact of the target parameter specification on variance and power. We compared the unadjusted estimator (i.e., the difference-in-means estimator) to the TMLE with various methods for initial estimation of the outcome regression $\mathbb{E}_0(Y|A, W)$. As predicted by theory, adjustment and pair-matching led to greater power. An estimator of the SATE was less variable than the same estimator of the PATE. While the differences in the estimators' variance were substantial, the differences in the attained power were attenuated due to the conservative variance estimator. Greater differences in the attained power were seen with a more aggressive fit of the outcome regression. As estimation of $\mathbb{E}_0(Y|A, W)$ improves, the TMLE becomes a more precise estimator (i.e., smaller true variance), and the variance estimator becomes less conservative. In small trials (e.g., $n \leqslant 30$) such as early phase clinical trials or cluster randomized trials, obtaining a precise estimate of $\mathbb{E}_0(Y|A, W)$ is likely to be challenging. In practice, many baseline covariates are predictive of the outcome, but adjusting for too many covariates can result in over-fitting. Ongoing work investigates the use of cross-validation in small trials to data adaptively select from a pre-specified library the optimal adjustment set [52]. As an area of future work, we plan to generalize these theorems and methods to observational studies. We hypothesize that a TMLE will provide at least as much precision and power to detect the impact of a non-randomized exposure on the study units (i.e., the SATE) than in some target population (i.e., the PATE).

Overall, we believe the sample effect is an interesting and possibly under-utilized causal parameter. It is simply the intervention effect for the study units. The SATE avoids assumptions about sampling from some vaguely defined target population. Furthermore, the SATE is responsive to heterogeneity in the treatment effect and avoids assumptions that the observed impact is generalizable or transportable to other contexts (e.g., [17–19]). These generalizations can be made with the formal methods and do not have to be assumed during the parameter specification. Furthermore, estimation of the SATE is likely to result in more precision and power to detect the exposure effect. To obtain a point estimate, the implementation of the TMLE is identical to that of the conditional and population estimands. To obtain conservative inference, we only need to take the sample variance of weighted residuals, divided by the appropriate sample size. Thereby, estimation and inference for the SATE does not require any extra work and is likely to give us more power to detect the impact of the exposure on the outcome.

Acknowledgements

Research reported in this publication was supported by Division of AIDS, NIAID of the National Institutes of Health under award numbers R01-AI074345, UM1AI069502, and U01AI099959 and in part by the President's Emergency Plan for AIDS Relief and Gilead Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, PEPFAR, or Gilead. The SEARCH project gratefully acknowledges the Ministries of Health of Uganda and Kenya, our research team, collaborators and advisory boards, and especially all communities and participants involved.

References

1. Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**:41–55.
2. Robins J. A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect. *Mathematical Modelling* 1986; **7**:1393–1512.
3. Horvitz D, Thompson D. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 1952; **47**:663–685.
4. Shen C, Li X, Li L. Inverse probability weighting for covariate adjustment in randomized studies. *Statistics in Medicine* 2014; **33**:555–568.
5. Snowden J, Rose S, Mortimer K. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *American Journal of Epidemiology* 2011; **173**(7):731–738.
6. Robins J, Rotnitzky A, Zhao L. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 1994; **89**:846–866.
7. van der Laan M, Robins J. *Unified Methods for Censored Longitudinal Data and Causality*. Springer-Verlag: New York Berlin Heidelberg, 2003.
8. van der Laan M, Rubin D. Targeted maximum likelihood learning. *The International Journal of Biostatistics* 2006; **2**(1):1–38. DOI: 10.2202/1557-4679.1043.
9. van der Laan M, Rose S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer: New York Dordrecht Heidelberg London, 2011.
10. Neyman J. 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* 1923; **5**:465–480.
11. Rubin DB. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science* 1990; **5**(4):472–480.
12. Imbens G. Nonparametric estimation of average treatment effects under exogeneity: a review. *Review of Economics and Statistics* 2004; **86**(1):4–29.
13. Imai K. Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Statistics in Medicine* 2008; **27**(24):4857–4873.

14. Schochet P. Estimators for clustered education RCTs using the Neyman model for causal inference. *Journal of Educational and Behavioral Statistics* 2013; **38**(3):219–238.
15. Imbens G, Rubin D. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press: New York, 2015.
16. Cole S, Stuart E. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 Trial. *American Journal of Epidemiology* 2010; **172**(1):107–115.
17. Stuart E, Cole S, Bradshaw C, Leaf P. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A* 2011; **174**(Part 2):369–386.
18. Bareinboim E, Pearl J. A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference* 2013; **1**(1):107–134.
19. Hartman E, Grieve R, Ramsahai R, Sekhon J. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society: Series A* 2015; **178**(3):757–778.
20. Balzer L, Petersen M, van der Laan M, the SEARCH Consortium. Adaptive pair-matching in randomized trials with unbiased and efficient effect estimation. *Statistics in Medicine* 2015; **34**(6):999–1011.
21. Klar N, Donner A. The merits of matching in community intervention trials: a cautionary tale. *Statistics in Medicine* 1997; **16**(15):1753–1764.
22. Greevy R, Lu B, Silber J, Rosenbaum P. Optimal multivariate matching before randomization. *Biostatistics* 2004; **5**(2): 263–275.
23. Imai K, King G, Nall C. The essential role of pair matching in cluster-randomized experiments, with application to the Mexican Universal Health Insurance Evaluation. *Statistical Science* 2009; **24**(1):29–53.
24. Hayes R, Moulton L. *Cluster Randomised Trials*. Chapman & Hall/CRC: Boca Raton, 2009.
25. van der Laan M, Balzer LB, Petersen M. Adaptive matching in randomized trials and observational studies. *Journal of Statistical Research* 2012; **46**(2):113–156.
26. Pearl J. Causal diagrams for empirical research. *Biometrika* 1995; **82**:669–710.
27. University of California, San Francisco. Sustainable East Africa Research in Community Health (SEARCH). ClinicalTrials.gov, 2013. Available from: http://clinicaltrials.gov/show/NCT01864603 [Accessed on 6 April 2016].
28. Pearl J. *Causality: Models, Reasoning and Inference*. Cambridge University Press: New York, 2000. Second ed., 2009.
29. van der Laan MJ, Hubbard AE, Kherad Pajouh S. Statistical inference for data adaptive target parameters, Technical Report 314, Division of Biostatistics, University of California, Berkeley. Available from: http://biostats.bepress.com/ucbbiostat/paper314/ [Accessed on 6 April 2016], 2013.
30. Hubbard A, van der Laan M. Mining with inference: data-adaptive target parameter. In *Handbook of Big Data*, P. Buhlmann, P. Drineas, M. Kane, M. van der Laan (eds). CRC Press, Taylor & Francis Group, LLC: Boca Raton, FL, 2016; 439–452.
31. Fisher RA. *Statistical Methods for Research Workers* (4th edn). Oliver and Boyd Ltd.: Edinburgh, 1932.
32. Cochran WG. Analysis of covariance: its nature and uses. *Biometrics* 1957; **13**:261–281.
33. Cox D, McCullagh P. Some aspects of analysis of covariance. *Biometrics* 1982; **38**(3):541–561.
34. Tsiatis A, Davidian M, Zhang M, Lu X. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statistics in Medicine* 2008; **27**(23):4658–4677.
35. Moore K, van der Laan M. Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. *Statistics in Medicine* 2009; **28**(1):39–64.
36. Rosenblum M, van der Laan M. Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. *The International Journal of Biostatistics* 2010; **6**(1):1–41. DOI: 10.2202/1557-4679.1138.
37. European Medicines Agency. Guideline on adjustment for baseline covariates in clinical trials. London, 2015.
38. Abadie A, Imbens G. Simple and bias-corrected matching estimators for average treatment effects, Technical Report 283, NBER Technical Working Paper, 2002.
39. Gruber S, van der Laan M. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *The International Journal of Biostatistics* 2010; **6**(1):1–14. DOI: 10.2202/1557-4679.1260.
40. Balzer L, Ahern J, Galea S, van der Laan M. Estimating effects with rare outcomes and high dimensional covariates: knowledge is power. *Epidemiologic Methods* In Press.
41. van der Laan M, Polley E, Hubbard A. Super learner. *Statistical Applications in Genetics and Molecular Biology* 2007; **6**(1):1–21.
42. Hahn J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 1998; **2**:315–331.
43. Gruber S, van der Laan M. TMLE: an R package for targeted maximum likelihood estimation. *Journal of Statistical Software* 2012; **51**(13):1–35.
44. Schwab J, Lendle S, Petersen M, van der Laan M, Gruber S. *LTMLE: longitudinal targeted maximum likelihood estimation*, 2015. Available from: http://CRAN.R-project.org/package=ltmle, R package version 0.9-6 [Accessed on 6 April 2016].
45. R Core Team. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2015. Available from: http://www.R-project.org [Accessed on 6 April 2016].
46. Fisher RA. *The Design of Experiments*. Oliver and Boyd Ltd.: London, 1935.
47. Zhang K, Small D. Comment: the essential role of pair matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation. *Statistical Science* 2009; **25**(1):59–64.

48. Lu B, Greevy R, Xu X, Beck C. Optimal nonbipartite matching and its statistical applications. *American Statistician* 2011; **65**(1):21–30.

49. Freedman L, Gail M, Green S, Corle D, The COMMIT Research Group. The efficiency of the matched-pairs design of the community intervention trial for smoking cessation (COMMIT). *Controlled Clinical Trials* 1997; **18**(2):131–139.

50. Campbell M, Donner A, Klar N. Developments in cluster randomized trials and *Statistics in Medicine*. *Statistics in Medicine* 2007; **26**(1):2–19.

51. Beck C, Lu B, Greevy R. *nbpMatching: functions for optimal non-bipartite optimal matching*, 2016. Available from: https://CRAN.R-project.org/package=nbpMatching, R package version 1.5.0 [Accessed on 6 April 2016].

52. Balzer L, van der Laan M, Petersen M. Adaptive pre-specification in randomized trials with and without pair-matching, Technical Report 336, Division of Biostatistics, University of California at Berkeley. Available from: http://biostats.bepress.com/ucbbiostat/paper336/ [Accessed on 6 April 2016], 2015.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.