Published in final edited form as:

Epidemiol Methods. 2016 December ; 5(1): 1–18. doi:10.1515/em-2014-0020.

Estimating Effects with Rare Outcomes and High Dimensional Covariates: Knowledge is Power

Laura Balzer,

Department of Biostatistics, Harvard T.H. Chan School of Public Health, 655 Huntington Ave, Boston, MA 02115, USA

Jennifer Ahern,

Department of Epidemiology, UC Berkeley, Berkeley, CA, USA

Sandro Galea, and

Department of Epidemiology, Boston University, Boston, MA, USA

Mark van der Laan

Department of Biostatistics, UC Berkeley, Berkeley, CA, USA

Abstract

Many of the secondary outcomes in observational studies and randomized trials are rare. Methods for estimating causal effects and associations with rare outcomes, however, are limited, and this represents a missed opportunity for investigation. In this article, we construct a new targeted minimum loss-based estimator (TMLE) for the effect or association of an exposure on a rare outcome. We focus on the causal risk difference and statistical models incorporating bounds on the conditional mean of the outcome, given the exposure and measured confounders. By construction, the proposed estimator constrains the predicted outcomes to respect this model knowledge. Theoretically, this bounding provides stability and power to estimate the exposure effect. In finite sample simulations, the proposed estimator performed as well, if not better, than alternative estimators, including a propensity score matching estimator, inverse probability of treatment weighted (IPTW) estimator, augmented-IPTW and the standard TMLE algorithm. The new estimator yielded consistent estimates if either the conditional mean outcome or the propensity score was consistently estimated. As a substitution estimator, TMLE guaranteed the point estimates were within the parameter range. We applied the estimator to investigate the association between permissive neighborhood drunkenness norms and alcohol use disorder. Our results highlight the potential for double robust, semiparametric efficient estimation with rare events and high dimensional covariates.

Keywords

bounded mean models; causal inference; rare outcomes; semiparametric estimation; targeted minimum loss-based estimation (TMLE)

1 Introduction

When the outcome of interest occurs infrequently, effect estimation can be particularly challenging. For example, a recent study sought to examine the impact of planned place of delivery (obstetric unit or not) on perinatal mortality and neonatal morbidities, occurring in 250 of 63,827 births (0.39 %) (Birth-place in England Collaborative Group 2011). Due to the paucity of individual birth events, however, the researchers estimated the effect on a composite outcome measure. Likewise, tuberculosis is a main cause of mortality among HIV + people (World Health Organization 2013). Evaluating strategies to reduce its transmission are essential, but difficult due to the disease's relatively low incidence. Along the same lines, international consortiums have been established to investigate the burden and treatment for uncommon cancers (e. g. RARECARENet 2014). While these outcomes are rare, a better understanding of their occurrence is likely to have important policy and health implications.

For binary outcomes or proportions, parametric logistic regression is often used to estimate the conditional odds ratio, given the exposure and measured covariates. Several researchers have investigated the performance of this approach when the outcome is extremely rare (e. g. Concato et al. 1993; Harrell et al. 1996; Peduzzi et al. 1996; King and Zeng 2001; Harrell 2001; Braitman and Rosenbaum 2002; Cepeda et al. 2003; Vittinghoff and McCulloch 2007). For example, simulations by Peduzzi et al. (1996) illustrated that estimates could be biased and inference unreliable if the number of outcomes per independent variable in the regression model was less than 10. The authors also found problems with estimator convergence, statistical power and the validity of significance tests (i. e. type I error rates and confidence interval coverage). Harrell et al. (1996) cautioned against over-fitting and encouraged the use of cross-validation or bootstrapping for model validation. Moreover, King and Zeng (2001) found that standard logistic models could substantially under-estimate the probability of the outcome and offered a bias correction with accompanying software.

When dealing with rare events, several researchers have recommended estimators based on the propensity score, which is the conditional probability of being exposed, given the covariates (Rosenbaum and Rubin 1983). These methods avoid estimation of the conditional mean outcome and thereby are expected to perform well when there are very few outcome events (e. g. Joffe and Rosenbaum 1999; Braitman and Rosenbaum 2002; Patorno et al. 2014). Simulations by Cepeda et al. (2003) suggested that propensity score methods were less biased and more efficient than logistic regression for the mean outcome, when the number of events per independent variable in the regression model was less than 8. The authors also cautioned that the performance of propensity score methods depended on the strength of the relationship between the covariates and the exposure.

Targeted minimum loss-based estimation (TMLE) is a general methodology for the construction of semiparametric, efficient substitution estimators (van der Laan and Rubin 2006; van der Laan and Rose 2011). A TMLE for a single time point exposure can be implemented as follows. First, the conditional expectation of the outcome, given the exposure and covariates, is estimated with parametric regression or with a more flexible approach, such as SuperLearner (van der Laan et al. 2007). Second, information on the exposure-covariate relation (i. e. the propensity score) is incorporated to improve this initial

estimator. The propensity score can also be estimated parametrically or with a more flexible approach. Informally, this “targeting” step helps remove some of the residual bias due to incomplete adjustment for confounding. More formally, this targeting step serves to solve the efficient score equation. Finally, the targeted predictions of the outcome under the exposure and under no exposure are averaged over the sample and contrasted on the relevant scale.

Thereby, TMLE requires estimation of both the conditional mean outcome as well as the propensity score, and achieves a number of desirable asymptotic properties (van der Laan and Rose; 2011). The standardized estimator is asymptotically normal with mean 0 and variance given by the variance of its influence curve. The TMLE is also double robust: if either the conditional mean outcome or the propensity score is consistently estimated, we will have a consistent estimate of the parameter of interest. If both functions are consistently estimated (at a fast enough rate), the TMLE will be efficient and achieve the lowest possible asymptotic variance among a large class of estimators. Finally, TMLE is a substitution estimator, providing robustness in the face of positivity violations (when there is no or little variability in the exposure within certain covariate strata) and rare outcomes (e. g. Stitelman and van der Laan 2010; Gruber and van der Laan 2010; Sekhon et al. 2011; Petersen et al. 2012; Gruber and van der Laan 2013; Lendle et al. 2013). Building on the work of Gruber and van der Laan (2010), this paper proposes a new TMLE for the semiparametric statistical model $\mathcal{M}$, which imposes bounds on the conditional mean of the outcome, given the exposure and measured confounders. We focus our discussion on rare binary outcomes and rare bounded continuous outcomes (e. g. proportions). The estimation problem and the theoretical motivation for the new TMLE are outlined in Section 2. In particular, the causal parameter, the corresponding statistical parameter, the statistical model and the efficient influence curve are discussed. Section 3 presents the rationale and procedure for the rare outcomes TMLE. Simulations and the applied analysis are given in the Section 4. The article concludes with a discussion of the advantages and disadvantages of the proposed method as well as areas for future work.

2 The estimation problem

We are interested in estimating the impact of a binary exposure A on the risk of a rare outcome Y . For example, Y might be an indicator that the subject develops tuberculosis with an incidence rate of 255/100,000 per person-year in Sub-Saharan Africa (World Health Organization 2013). Alternatively, Y might be the one-year cumulative incidence of tuberculosis for a given community. In the latter scenario, the outcome Y is a proportion bounded between $[0, 1]$. Suppose we measure some baseline characteristics W that are predictors of both the exposure and outcome. In other words, W represents the set of measured confounders. Let $O = (W, A, Y)$ denote the observed data random variable with distribution P_0 . Throughout, subscript 0 denotes the true, but unknown distribution.

To translate our scientific question into a causal parameter, let us define Y_a as the counterfactual outcome, if possibly contrary-to-fact, the unit received exposure $A = a$ (Neyman 1923; Rubin 1974; Pearl 2000). We assume these quantities exist for all units both under the exposure ($A = 1$) and under no exposure ($A = 0$). To relate the observed outcomes

to the counterfactual outcomes, we need the stable unit treatment value assumption (SUTVA) (Rubin 1978; 1980): (1) the counterfactual outcomes for one unit must not be impacted by the treatment assignment of another unit (i. e. no interference), and (2) there must not be multiple versions of the treatment $A = a$. Under this assumption, we have

$$Y = AY_1 + (1 - A)Y_0.$$

In words, we only observe the counterfactual outcome corresponding to the observed treatment $Y_A = Y$. Thereby, the observed data can be considered as a time-ordered missing data structure on the full data $X^F = (W, Y_1, Y_0) \sim P_X$, with the exposure A as the censoring variable (Neyman 1923; Rubin 1974). Throughout our goal is to estimate and obtain inference for the population average treatment effect:

$$\Psi^F(P_X) = E(Y_1) - E(Y_0).$$

This causal parameter is the difference in expected counterfactual outcomes if everyone in the target population were exposed and if everyone in that target population were unexposed. For a binary outcome, $\Psi^F(P_X)$ corresponds to the causal risk difference: $P(Y_1 = 1) - P(Y_0 = 1)$.

To express $\Psi^F(P_X)$ as a function of the observed data distribution P_0 , we need several assumptions. First, there must be no unmeasured confounders of the effect of the exposure on the outcome (Rosenbaum and Rubin 1983; Robins 1986). Secondly, there must be sufficient variability in the treatment assignment. In other words, the propensity score $P_0(A = 1 | W)$ must be bounded away from 0 and 1. This condition is known as the positivity assumption. Under these assumptions, we can express the causal parameter in terms of the difference in the conditional mean outcomes, averaged (standardized) with respect to the covariate distribution (Robins 1986):

$$\Psi(P_0) = \sum_w [E_0(Y | A=1, W=w) - E_0(Y | A=0, W=w)] P_0(W=w) = E_0[\bar{Q}_0(1, W) - \bar{Q}_0(0, W)]$$

where the summation generalizes to the integral for continuous covariates and $\bar{Q}_0(A, W) = E_0(Y | A, W)$ denotes the conditional mean outcome, given the exposure and covariates. For a binary outcome $\Psi(P_0)$ is sometimes called the marginal risk difference.

The challenge, addressed in this paper, is estimation of $\Psi(P_0)$ in the context of extremely rare outcomes and high dimensional covariates W . This challenge is illuminated by studying the efficient influence curve (function) of the target parameter Ψ at the true probability distribution P_0 (Bickel et al. 1993; van der Laan and Rose 2011):

$$D^*(P_0)(O) = \left(\frac{\mathbb{I}(A=1)}{P_0(A=1|W)} - \frac{\mathbb{I}(A=0)}{P_0(A=0|W)} \right) (Y - \bar{Q}_0(A, W)) + \bar{Q}_0(1, W) - \bar{Q}_0(0, W) - \Psi(P_0) \quad (1)$$

where $\mathbb{I}(\cdot)$ is the indicator function. $D^*(P_0)(O)$ is the critical ingredient in the construction of double-robust, semiparametric, efficient estimators (van der Laan and Robins 2003; van der Laan and Rose 2011). Specifically, $D^*(P_0)(O)$ can be represented as an estimating function in $\Psi(P_0)$ as in augmented inverse probability of treatment weighting (A-IPTW) or used to fluctuate initial estimates in the substitution estimator TMLE.

The information for learning the target parameter is captured by the sample size n divided by the variance of the efficient influence curve at the true distribution P_0 :

$$\text{Information} = \frac{n}{\text{Var}[D^*(P_0)(O)]} \quad (2)$$

(van de Vaart 1998). Low information can occur when the propensity score $P_0(A = 1|W)$ is or approaches 0 or 1 for certain treatment-covariate combinations. The impact of violations of the positivity assumption on estimator performance are demonstrated in Petersen et al. (2012), among others. Low information can also occur when there are very few outcome events (i. e. $E_0(Y)$ is small) and the conditional mean $\bar{Q}_0(A, W)$ is or approaches 1 for some treatment-covariate combinations. In either case, the variance of the efficient influence curve can become large, and the sample size needed to detect effects with sufficient power correspondingly large. In this paper, we address estimation in the context of low information due to rare outcomes and high dimensional covariates.

We consider the semiparametric statistical model $\mathcal{M}$, which incorporates knowledge that the conditional mean of the rare outcome $\bar{Q}_0(A, W)$ is bounded from above by some $u < 1$. For rare binary outcomes, this knowledge has been noted by other researchers (Beck, King and Zeng, 2000, King and Zeng, 2001). Consider, for example, an unusual outcome with incidence 0.5 %. With this rare of an event, it seems unlikely that the conditional mean outcome exceeds 10% for any combination of the exposure and measured confounders. Instead, the researcher might have knowledge that this conditional mean does not exceed 5% for the study population. Formally, the statistical model $\mathcal{M}$ is the set of possible data generating distributions:

$$\mathcal{M} = \{P: \bar{Q}(A, W) \in [\ell, u]\}$$

for some bounds $0 < \ell < u < 1$. The specification of the bounds can be based on subject matter knowledge or selected with cross-validation, as discussed in Section 3.3. We place no

restrictions on the marginal distribution of baseline covariates $P(W)$ or on the propensity score $P(A = 1|W)$.

The variance of the efficient influence curve (Eq. [1]) and the information for estimating the target parameter (Eq. [2]) depend on whether the true distribution P_0 and thereby the true conditional mean $\bar{Q}_0(A, W)$ satisfy the constraints in the semiparametric model $\mathcal{M}$. Likewise, the asymptotic variance of a TMLE (under sampling from P_0) depends on whether the limit of the estimator of $\bar{Q}_0(A, W)$ satisfies these constraints. Specifically, the variance of the efficient influence curve and the variance of the influence curve of the TMLE will be smaller when the true conditional mean and the limit of the estimator are in $[\ell, u]$ for some $u < 1$ (Appendix A). Furthermore, when the outcome is rare and the dimension of the covariates large, we are often forced to use a misspecified estimator for $\bar{Q}_0(A, W)$. Then the influence curve of the TMLE is defined by the efficient influence curve at the misspecified limit $\bar{Q}(A, W)$. The constraints in $\mathcal{M}$ heavily affect the variance of this influence curve and therefore motivate us to develop an estimator that respects the known bounds (Appendix B).

In the following section, we propose a new TMLE, which guarantees that the initial and targeted estimator of $\bar{Q}_0(A, W)$ are constrained within the model bounds. We compare the proposed modification to the standard TMLE algorithm for binary or bounded continuous outcomes (Gruber and van der Laan 2010). The two TMLEs will be asymptotically equivalent if both use a consistent estimator of the conditional mean outcome $\bar{Q}_0(A, W)$ or more generally if their estimators of $\bar{Q}_0(A, W)$ converge to the same limit. In finite samples, however, enforcing the constraints is expected to have substantial impact on performance. Indeed, the finite sample behavior of a TMLE is largely driven by the sample variance of the influence curve at the targeted estimator $\bar{Q}_n^*(A, W)$, and this variance is expected to be much smaller when $\bar{Q}_n^*(A, W)$ respects the model constraints. As a result, we expect and show in simulations the proposed method will result in tighter confidence intervals and more power in finite samples.

Before introducing the new TMLE, we first caution that the statistical model $\mathcal{M}$ must contain the true observed data distribution P_0 . In other words, the assumed bounds on the conditional mean outcome $\bar{Q}_0(A, W)$ should be given careful consideration and not chosen based on convenience. For example, there might be a scenario, where the marginal expectation of the outcome Y is small (e. g. $E_0(Y) < .01$), but the conditional mean $\bar{Q}_0(A, W)$ approaches 1 for some combinations of the exposure and measured confounders. Consider, for example, a study of the association between very preterm birth (28 to < 32 weeks) and neonatal mortality. If the set of measured confounders included socio-economic factors, race, maternal smoking and maternal anemia, then the conditional risk of the outcome could approach 1 for some infants. As a result, it would be inappropriate to assume a semiparametric statistical model $\mathcal{M}$ with upper bound $u < 1$. Enforcing a small upper bound may result in bias and lower power in finite samples. Instead, for this example, an upper bound of $u = 1$ would be appropriate, and the new estimator, proposed in the following section, would reduce to the standard TMLE procedure for binary or bounded continuous outcome. For the remainder of the article, we assume that the semiparametric statistical

model $\mathcal{M}$ characterizes the set of possible observed data distributions in that the true conditional mean outcome $\bar{Q}_0(A, W)$ is bounded in $[\ell, u]$ for some $0 \leq \ell < u < 1$.

3 Estimating effects with rare outcomes

Suppose we have n observations $O = (W, A, Y)$ drawn independently from P_0 . A simple substitution estimator (i. e. the G-Computation estimator of Robins (1986)) for $\Psi(P_0)$ can be implemented with the following steps. First, we obtain an estimate of the conditional mean outcome $\bar{Q}_0(A, W)$. Let us denote an initial estimator based on n observations as $\bar{Q}_n(A, W)$. Then we obtain the predicted outcomes for each observation under the exposure $\bar{Q}_n(1, W)$ and under the control $\bar{Q}_n(0, W)$. The sample average difference in these predicted outcomes provides a point estimate for $\Psi(P_0)$. The final step corresponds to estimating the marginal covariate distribution with the empirical proportion.

In some cases, we can estimate $\bar{Q}_0(A, W)$ with the sample average in each exposure-covariate strata. This non-parametric estimator can quickly become ill-defined when there are many exposure-covariate combinations and can suffer from over-fitting, especially with rare outcomes. In some cases, we may have the background knowledge to support parametric regression for $\bar{Q}_0(A, W)$. Suppose, for example, the true conditional mean outcome could be described by the following parametric regression model.

$$\bar{Q}_0(A, W) = \text{logit}^{-1}[\beta_0 + \beta_1 A + \beta_2 W_1 + \dots + \beta_{(p+1)} W_p] \in [\ell, u]$$

where $\text{logit}(x) = \log(x/1-x)$ and $W = (W_1, \dots, W_p)$ denotes the set of confounders. Asymptotically, the estimated coefficients will yield predictions of the mean outcome, which respect the model knowledge, and estimates of the target parameter, which are consistent and efficient. However, with small sample sizes and many confounders, estimation will rely heavily on extrapolation and may yield estimates outside of the model bounds. In this realistic situation, incorporating knowledge of the upper bound u into the estimation algorithm is expected to improve performance. In particular, we conjecture that even if the parametric regression is misspecified, then using the bound u will result in more precision and less bias. Nonetheless, we rarely have the knowledge to specify the form of $\bar{Q}_0(A, W)$ with a finite number of parameters (i. e. β). As a result, we advocate for a more flexible approach that respects our semiparametric statistical model $\mathcal{M}$.

Data-adaptive or machine learning algorithms can incorporate our limited model knowledge, while smoothing over areas of data with weak support. SuperLearner, for example, allows a set of pre-specified algorithms to compete and selects the best algorithm using cross-validation (i. e. data-splitting) (van der Laan et al. 2007). The selected algorithm can be used to predict the outcomes for all units given the exposure and their measured covariates $\bar{Q}_n(1, W)$ as well as given no exposure and their measured covariates $\bar{Q}_n(0, W)$. The average difference in these predicted outcomes provides a point estimate of the target parameter. Inference, however, must respect the model building process. (Treating the selected algorithm as if it were *a priori*-specified can result in misleading inference.) Furthermore, estimating the conditional expectation $\bar{Q}_0(A, W)$ is a more ambitious task than estimating a

single number $\Psi(P_0) \in \mathbb{R}$. Thereby, an estimator of $\bar{Q}_0(A, W)$ will have the wrong bias-variance trade-off for the parameter of interest $\Psi(P_0)$. Specifically, the estimator will have too much bias relative to its variance.

TMLE provides a solution to several of these challenges (van der Laan and Rose 2011). TMLE incorporates data-adaptive estimation with an additional targeting step to reduce bias for $\Psi(P_0)$ and to attain valid statistical inference. As described in the introduction, a TMLE for $\Psi(P_0)$ can be implemented in three steps. First, we estimate of the conditional mean outcome $\bar{Q}_0(A, W)$ with parametric or more flexible methods. Second, we target this initial estimator using information in the estimated propensity score $P_n(A = 1|W)$. Finally, we substitute the targeted estimates into the parameter mapping. The updating step is accomplished with a loss function and a carefully selected submodel. For a binary or bounded continuous outcome, we can use the negative log-likelihood loss function (Wedderburn 1974; McCullagh 1983; Gruber and van der Laan 2010):

$$-\mathcal{L}(\bar{Q})(O) = \log \left[\bar{Q}(A, W)^Y (1 - \bar{Q}(A, W))^{1-Y} \right], \quad (3)$$

and the following logistic regression model to update the initial estimator:

$$\text{logit}[\bar{Q}_n(A, W)(\varepsilon)] = \text{logit}[\bar{Q}_n(A, W)] + \varepsilon H_n(A, W)$$

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

Here, ε is the univariate parameter and $P_n(A = 1|W)$ denotes an estimate of the propensity score. This combination of loss function and fluctuation model returns the initial estimator $\bar{Q}_n(A, W)$ at $\varepsilon = 0$ and has score equal to the relevant component of the efficient influence curve at $\varepsilon = 0$ (Gruber and van der Laan 2010). To estimate the fluctuation parameter ε , we run logistic regression of the outcome Y on the covariate $H_n(A, W)$ with the *logit* of the initial estimates as offset. Plugging in the estimated coefficient ε_n yields the targeted updates. A point estimate can then be obtained by taking the average difference in the targeted predictions under the exposure $\bar{Q}_n^*(1, W)$ and under no exposure $\bar{Q}_n^*(0, W)$.

The standard TMLE algorithm for a binary or bounded continuous outcome is not optimal for our semiparametric statistical model $\mathcal{M}$. Specifically, this TMLE does not enforce the global constraints on the conditional mean outcome. First, initial estimates $\bar{Q}_n(A, W)$ based on the log-likelihood loss function (Eq. [3]) are guaranteed to be within $(0, 1)$, but are possibly outside of the model bounds $[\ell, u]$. Second, the logistic regression, used to update the initial estimator, also does not respect the constraints implied by $\mathcal{M}$. Therefore, the updated estimates $\bar{Q}_n^*(A, W)$ are guaranteed to be between $(0, 1)$, but may be outside of the model bounds $[\ell, u]$. As a result, the asymptotic and finite sample performance of a TMLE, using this loss function and parametric working model, is expected to be sub-optimal. The

algorithm is expected to be unstable (i. e. more variable and less accurate) when the outcome is very rare and there are many confounders.

3.1 The rare outcomes TMLE

To incorporate model knowledge, we propose a linear transformation of the rare outcome $Y \in [0, 1]$ by subtracting off the lower bound ℓ and dividing by the deviation between the upper u and lower bounds:

$$\tilde{Y} = \frac{Y - \ell}{u - \ell} \in \left[\frac{-\ell}{u - \ell}, \frac{1 - \ell}{u - \ell} \right].$$

The analogous transformation was proposed by Gruber and van der Laan (2010) for the TMLE of a bounded continuous outcome $Y \in [a, b]$. Suppose, for example, the bounds on the conditional mean outcome $\bar{Q}_0(A, W)$ are $[0, 0.05]$. Then the transformed outcome $\tilde{Y}$ would be bounded between 0 and 20. The mapping between the conditional mean of the original outcome Y and the conditional mean of the transformed outcome $\tilde{Y}$ is given by

$$\bar{Q}(A, W) = \ell + (u - \ell)\tilde{Q}(A, W).$$

Therefore, the conditional mean of the transformed outcome $\tilde{Y}$ remains bounded in $[0, 1]$:

$$\tilde{Q}(A, W) = \frac{\bar{Q}(A, W) - \ell}{u - \ell} \in \left[\frac{\ell - \ell}{u - \ell}, \frac{u - \ell}{u - \ell} \right] = [0, 1].$$

As a result, the negative quasi-log-likelihood (Wedderburn 1974; McCullagh 1983) is a valid loss function for initial estimation and targeting of the transformed mean $\tilde{Q}_0(A, W)$:

$$-\mathcal{L}(\tilde{Q})(O) = \log \left[\tilde{Q}(A, W)^{\tilde{Y}} (1 - \tilde{Q}(A, W))^{1 - \tilde{Y}} \right].$$

The proof is analogous to Lemma 1 in Gruber and van der Laan (2010) and thus omitted here.

To update an initial estimator of $\tilde{Q}_n(A, W)$, we can use the logistic fluctuation submodel: $\text{logit}[\tilde{Q}_n(A, W)(\epsilon)] = \text{logit}[\tilde{Q}_n(A, W)] + \epsilon H_n(A, W)$ with covariate $H_n(A, W)$ defined as above. This combination of loss function and fluctuation submodel will generate a score proportional to the relevant component of the efficient influence curve at zero fluctuation:

$$\begin{aligned} \frac{d}{d\epsilon} \mathcal{L}(\tilde{Q}_n(\epsilon))(O)|_{\epsilon=0} &= H_n(A, W)(\tilde{Y} - \tilde{Q}_n(A, W)) \\ &= H_n(A, W) \left(\frac{Y - \ell}{u - \ell} - \tilde{Q}_n(A, W) \right) \\ &= \frac{1}{u - \ell} H_n(A, W)(Y - \bar{Q}_n(A, W)). \end{aligned}$$

Through this transformation, the initial and targeted estimates are guaranteed to satisfy the model constraints. This will provide robustness. The targeted estimates can then be rescaled and substituted in the parameter mapping. The proposed loss function and parametric submodel define a new TMLE of the target parameter $\Psi(P_0)$ in the semiparametric statistical model $\mathcal{M}$, which encodes bounds on the conditional mean of the rare outcome $\bar{Q}_0(A, W) \in [\ell, u]$. Hereafter, we refer to the proposed estimator as the rare outcomes TMLE.

3.2 Step-by-step implementation

We present the rare outcomes TMLE for the semiparametric model $\mathcal{M}$, containing knowledge on both the lower bound $0 \leq \ell$ and upper bound $u \leq 1$ on the conditional mean outcome. In the context of rare outcomes, the lower bound will often be set to $\ell = 0$.

Step 1: Transform the outcome—We first transform the outcome Y into $\tilde{Y}$ by subtracting the lower bound ℓ and dividing by the difference between the upper and lower bounds $(u - \ell)$.

Step 2: Estimate the transformed mean—An initial estimate of the conditional mean of the transformed outcome $\tilde{Q}_0(A, W)$ can be based on logistic regression of $\tilde{Y}$ on the exposure A and baseline covariates W . Since the outcome is no longer between 0 and 1, standard software may yield error messages. Example R code, using the `optim` function, is given in the Appendix (R Core Team 2014). More data-adaptive methods, such as SuperLearner, can be implemented as long as the library of algorithms respect the statistical model (van der Laan et al. 2007).

Step 3: Estimate the propensity score—An initial estimate of the conditional probability of being exposed given the covariates $P_0(A = 1 | W)$ is also required, as it makes up the covariate $H_n(A, W)$ in the fluctuation submodel. Using the log-likelihood loss function, the propensity score could be estimated with parametric logistic regression or with more data-adaptive methods.

Step 4: Target the initial estimator—Run logistic regression of the transformed outcome $\tilde{Y}$ on the covariate $H_n(A, W)$ with offset as the *logit* of the initial estimates $\tilde{Q}_n(A, W)$. The estimated coefficient ε_n is then plugged into yield the updates:

$$\tilde{Q}_n^*(A, W) = \tilde{Q}_n(A, W) + \varepsilon_n H(A, W).$$

The process of estimating the fluctuation parameter ε and updating is iterated until convergence, which occurs here in a single step.

Step 5: Transform and plug-in the targeted estimates—The targeted estimates of the conditional mean of the transformed outcome, denoted $\tilde{Q}_n^*(A, W)$, can be mapped into targeted estimates of the conditional mean of the original outcome by

$$\bar{Q}_n^*(A, W) = \ell + (u - \ell) \tilde{Q}_n^*(A, W).$$

We obtain a point estimate by substituting in the targeted estimates of the conditional mean under the exposure $\bar{Q}_n^*(1, W)$ and under no exposure $\bar{Q}_n^*(0, W)$ into the parameter mapping:

$$\Psi_{n,rtmle}(P_n) = \frac{1}{n} \sum_{i=1}^n (\bar{Q}_n^*(1, W_i) - \bar{Q}_n^*(0, W_i)).$$

Step 6: Obtain inference—Under regularity conditions, $\Psi_{n,rtmle}(P_n)$ is an asymptotically linear estimator of $\Psi(P_0)$ (van der Laan and Rose 2011). Its limit distribution is normal with mean 0 and variance given by the variance of its influence curve divided by sample size n . Therefore, 95% confidence intervals can be constructed as $\psi_n^* \pm 1.96\sigma_n/\sqrt{n}$, where ψ_n^* denotes the resulting point estimate and where σ_n^2 is the sample variance of the estimated influence curve:

$$IC_n(O) = \left(\frac{\mathbb{I}(A=1)}{P_n(A=1|W)} - \frac{\mathbb{I}(A=0)}{P_n(A=0|W)} \right) (Y - \bar{Q}_n^*(A, W)) + \bar{Q}_n^*(1, W) - \bar{Q}_n^*(0, W) - \psi_n^*.$$

Alternative approaches for variance estimation include the non-parametric bootstrap or a substitution estimator for the variance. The former might be problematic with rare binary outcomes as some bootstrapped samples may not have any events. The latter would guarantee the bounds on the variance are respected and is an area of future work.

By construction, the rare outcomes TMLE solves the efficient score equation. Specifically, the empirical mean of the efficient influence curve at the targeted estimator $\bar{Q}_n^*(A, W)$ and initial estimator $P_n(A=1|W)$ is zero. As a result, the TMLE is double robust: ψ_n^* will be a consistent estimator for $\psi_0 = \Psi(P_0)$ if either the conditional mean function $\bar{Q}_0(A, W)$ or the propensity score $P_0(A=1|W)$ is consistently estimated. As shown below, this property translates into important bias gains in finite samples. If both functions are consistently estimated at a fast enough rate and the propensity score satisfies the positivity assumption, the proposed TMLE will be asymptotically efficient in that its influence curve equals the efficient influence curve. As illustrated below, this property translates into important variance and power gains in finite samples.

3.3 Selecting the upper bound u with cross-validation

Thus far, we have assumed that the upper bound u is known. (For rare outcomes, the lower bound l can trivially be set to 0.) Such knowledge, however, may be unavailable in all applications. When the specified upper bound is larger than needed, the gains in estimator performance will be attenuated. In the extreme when $u = 1$, the rare outcomes TMLE will reduce to the standard TMLE algorithm for binary or bounded continuous outcomes. If the specified upper bound is too small, then the targeted estimator of $\bar{Q}_0(A, W)$ will be inconsistent. Nonetheless, if the propensity score is known or consistently estimated, then the target parameter ψ_0 will still be consistently estimated due to the double robustness property. Moreover, the targeted estimates $\bar{Q}_n^*(A, W)$ will still be bounded from above by u , which translates into important variance gains. Thereby, cross-validation can be used to

select u , when such knowledge is not available *a priori*. Specifically, the upper bound can be selected by minimizing the cross-validated risk of candidate estimators $\bar{Q}_{n,u}(A, W)$, which are now indexed by an upper bound u . The previously stated properties (e. g. double robustness, asymptotic linearity and efficiency) should hold when the upper bound u is selected data-adaptively. We implement this cross-validation selector in the following simulations.

4 Simulations & data application

The following simulation studies compare the finite sample performance of the standard TMLE with the proposed rare outcomes TMLE (rTMLE). For comparison, we also include a propensity score matching (PSM) estimator (Rosenbaum and Rubin 1983), inverse probability of treatment weighted (IPTW) estimator (Hernán et al. 2000) and augmented inverse probability of treatment weighted (AIPTW) estimator (Robins 2000; van der Laan and Robins 2003). The first two methods rely solely on estimation of the propensity score and may have superior performance with very rare outcomes. AIPTW requires estimation of both the conditional mean outcome and the propensity score, but is double robust and asymptotically efficient under consistent estimation of both functions. AIPTW is an estimating equation (i. e. not a substitution estimator) and thereby can result in impossible parameter estimates (e. g. probabilities less than 0 or greater than 1) (Lendle et al. 2013).

For the PSM estimator, we used the `Matching` package (Sekhon 2011) for 1:1 matching based on the estimated propensity score and calculated the point estimate as

$$\psi_{n,PSM} = \frac{1}{n} \sum_{i=1}^n \hat{Y}_i(1) - \hat{Y}_i(0), \text{ where } \hat{Y}_i(a) = \begin{cases} Y_i & \text{if } A_i = a \\ Y_{M_i} & \text{if } A_i \neq a \end{cases}$$

with Y_{M_i} denoting the outcome of the observation matched to unit i based on the estimated propensity scores. A point estimate from IPTW is given by the following weighted mean

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

A point estimate from AIPTW is attained by directly solving the efficient score equation:

$$\psi_{n,AIPTW} = \frac{1}{n} \sum_{i=1}^n \left[\left(\frac{\mathbb{I}(A_i=1)}{P_n(A_i=1|W)} - \frac{\mathbb{I}(A_i=0)}{P_n(A_i=0|W)} \right) (Y_i - \bar{Q}_n(A_i, W_i)) + \bar{Q}_n(1, W_i) - \bar{Q}_n(0, W_i) \right]$$

where $\bar{Q}_n(A, W)$ denotes a non-targeted estimate of the conditional mean outcome. Inference was based on the Abadie-Imbens standard error estimator for the PSM estimator (Abadie and Imbens 2006; Sekhon 2011) and the estimated influence curve for the others. Wald-type confidence intervals were calculated as $\psi_n \pm 1.96se_n$, where ψ_n denotes the point estimate and se_n denotes the estimated standard error. Likewise, tests of the null hypothesis of no effect were based on test statistic $T_n = \psi_n/se_n$.

4.1 Simulation 1: individual-level data

The finite sample performance of the estimators was assessed by drawing 2,000 samples of sizes $n = 1000$ and $n = 2500$ according to the following process. First, we generated three baseline covariates:

$$W1 \sim \text{Normal}(0, 0.25^2), \quad W2 \sim \text{Uniform}(0, 1), \quad W3 \sim \text{Bernoulli}(0.5).$$

The exposure A was drawn from a Bernoulli distribution with probability

$$P_0(A=1|W) = \text{logit}^{-1}[-0.5 + W1 + W2 + W3].$$

With this exposure mechanism, there were no positivity violations; the propensity score was bounded between 20% and 92 %. Finally, the binary outcome was drawn from a Bernoulli distribution with probability

$$P_0(Y=1|A, W) = \text{logit}^{-1}[-3 + 2 \cdot A + W1 + 2 \cdot W2 - 4 \cdot W3 + 0.5 \cdot A \cdot W1]/15.$$

The resulting marginal probability of the outcome was $E_0(Y) = 1.1\%$. The true bounds on the conditional mean $\bar{Q}_0(A, W)$ were $[0\%, 6.2\%]$, and the true value of the statistical parameter was $\psi_0 = 1.3\%$. The propensity score $P_0(A=1|W)$ was estimated with the correctly specified main terms logistic model and a misspecified regression model failing to adjust for $W3$. The conditional mean outcome $\bar{Q}_0(A, W)$ was estimated with the correctly specified logistic regression model as well as a misspecified regression with main terms for only A and $W1$. For the rare outcomes TMLE, the lower bound was set 0 and the upper bound was selected data-adaptively from the set $\{2.5\%, 5\%, 7.5\%, 10\%\}$ using cross-validation with the log-likelihood loss function.

The results for this simulation are presented in Figures 1 and 2. As expected, the PSM estimator and IPTW exhibited low bias when the regression model for the propensity score $P_0(A=1|W)$ was correctly specified, but were biased otherwise. This bias did not disappear with sample size. Likewise, these estimators had poor confidence interval coverage under misspecification of the propensity score. Even under correct specification of this function, the PSM estimator had notably lower power than IPTW, AIPTW or rTMLE. Indeed, neither this PSM estimator nor this IPTW are efficient estimators. (The reader is referred to Abadie and Imbens (2015) for recent work on consistent estimation of the standard error when estimating the propensity score.)

In theory, the other estimators solve the efficient score equation directly (AIPTW) or during the targeting step (TMLE, rTMLE). As a result, these estimators are double robust. For these simulations, however, the standard TMLE exhibited substantial bias when estimating $\bar{Q}_0(A, W)$ with the correctly specified regression model, which included four main terms plus an interaction (Figure 1). Due to the paucity of outcome events, logistic regression of the untransformed outcome Y was unstable. Thereby, the updating step, which involved fitting an additional coefficient ϵ , did little to reduce bias. As a result, its confidence interval

coverage was poor under correct specification of the outcome regression (Figure 2). On the other hand, when the initial estimator of $\bar{Q}_0(A, W)$ was based on the misspecified regression, the performance of the standard TMLE was comparable to AIPTW and rTMLE. Recall the misspecified regression model for $\bar{Q}_0(A, W)$ only included main terms for the exposure A and confounder $W1$. The smaller adjustment set led to more stable initial estimates, but also complete reliance on consistent estimation of the propensity score for confounding control and consistency. Thereby, a potential approach when estimating effects with rare outcomes is to use a smaller adjustment set for $\bar{Q}_0(A, W)$ and thereby sacrifice double robustness and efficiency for greater stability and potentially reduced bias. This approach was taken by Gruber and van der Laan (2013).

The performance of AIPTW and rTMLE did not suffer when fitting $\bar{Q}_0(A, W)$ in the larger, correctly specified regression model. As shown in Figure 1, both AIPTW and rTMLE had low bias when either the outcome regression, the propensity score or both were correctly specified. Furthermore, these estimators had good confidence interval coverage at both sample sizes and attained power for the larger sample (Figure 2). Therefore, it seemed that the performance of AIPTW and rTMLE were comparable for this set of simulations. It is worth emphasizing, however, that rTMLE is a substitution estimator and thereby guaranteed to respect the global constraints in the model. In particular, AIPTW yielded negative (impossible) estimates for the marginal risk under no exposure $E_0[E_0(Y|A=0, W)]$ (Table 3 in the Appendix).

Under the null, the PSM estimator and IPTW had good type I error control when the propensity was correctly specified (Results not shown). Under misspecification, their type I error rates exceeded 20 %. The standard TMLE also suffered from high type I error rates, when both regression models were correctly specified. AIPTW and rTMLE maintained nominal type I error rates for both sample sizes and all regression specifications.

4.2 Simulation 2: group-level data

For this set of simulations, we focused on clustered data, where the covariates, exposure and outcome were measured or aggregated to the cluster-level. For 2,000 simulations of $n = 100$ units, three baseline covariates ($W1, W2, W3$) were drawn from a multivariate normal with means 0, standard deviation 0.25 and correlation 0.5. We generated two additional covariates ($W4, W5$) by drawing from a uniform distribution and a Bernoulli distribution with probability 0.5, respectively. The exposure A was drawn from a Bernoulli distribution with probability

$$P_0(A=1|W)=\text{logit}^{-1}[-0.5+W1+W4+W5].$$

With this exposure mechanism, there were no positivity violations; the propensity score was bounded between 22% and 92 %. The cluster-level outcome Y was the empirical mean of 2,500 independent Bernoulli's with a cluster-specific risk of

$$P_0(Y=1|A, W)=\text{logit}^{-1}[-1.5+0.35*A+0.75*W1-W2+0.75*W3+W4-W5+0.2*A*W1-0.2*A*W5+0.85*U_Y]/15$$

where U_Y was drawn from a normal distribution with mean 0 and standard deviation 0.25. The average outcome across the clusters was $E_0(Y) = 1.5\%$. The true bounds on the conditional mean $\tilde{Q}_0(A, W)$ were $[0\%, 5.2\%]$, and the true value of the statistical parameter was $\psi_0 = 0.30\%$. For completeness, the null scenario was also generated by simulating the outcomes as if all clusters were exposed.

As suggested by the previous simulations, a possible approach for estimation with rare outcomes is using a small adjustment set for $\tilde{Q}_0(A, W)$ and fully relying on consistent estimation of the propensity score $P_0(A = 1|W)$. To evaluate this approach, we estimated the conditional mean outcome with the unadjusted treatment-specific mean $\tilde{Q}_n(A)$. For comparison, we also employed SuperLearner for initial estimation of the transformed mean $\tilde{Q}_0(A, W)$ in rTMLE (Polley and van der Laan 2013). For simplicity, we limited our library of algorithms to logistic regressions, building up from a single main term to four terms (the exposure plus three covariates). For example, a candidate algorithm included as main terms the exposure A , the baseline covariates $W1$, $W4$ and $W5$. This simple library was selected for illustration; in practice, the inclusion of more flexible algorithms is recommended. The bounds for rTMLE were selected from the set $\{2.5\%, 5\%, 7.5\%, 10\%\}$ with cross-validation and the log-likelihood loss. The propensity score was estimated according to the correctly specified logistic regression model.

The simulation results are given in Table 1. Since the estimated propensity score was based on the correct regression model, all algorithms exhibited low bias. When using an unadjusted (initial) estimator $\tilde{Q}_n(A)$, the double robust estimators did not substantially outperform the PSM estimator. Instead, AIPTW, TMLE and rTMLE were under-powered, and there was evidence of conservative inference, as indicated by over-coverage of confidence intervals and lower than nominal Type I error rates. The rTMLE, using SuperLearner for initial estimation, had near nominal confidence interval coverage and type I error control. Furthermore, when there was a treatment effect, the rTMLE, using SuperLearner, attained highest power of 94 %, while the attained power of the other estimators was 40 %. This demonstrates the potential gains with data-adaptive estimation in the rTMLE algorithm.

4.3 Application

For a practical demonstration, we used data from the New York Social Environment Study to examine the association between permissive neighborhood drunkenness norms and the prevalence of alcohol use disorder at the neighborhood level ($n=59$). The study and resulting data structure are described in Ahern et al. (2008). Briefly, 4,000 telephone interviews were conducted among New York City residents to examine neighborhood exposures and mental health and substance use outcomes in 2005. The exposure of interest A was an indicator that the neighborhood had permissive drunkenness norms. The outcome Y was the proportion of neighborhood respondents meeting the DSM-5 criteria of alcohol use disorder within a 12-month period. This outcome combines the previous categories of alcohol abuse and alcohol dependence (National Institute on Alcohol Abuse and Alcoholism 2013), and its low prevalence (median of 1.5% across neighborhoods) had previously prohibited its examination. The set of measured confounders W included age, race, gender, marital status,

birthplace, interview language, education, income level, employment status, years lived in neighborhood and history of drinking. The parameter of interest was the marginal risk difference $\Psi(P_0)$, which can be interpreted as the difference in the strata-specific risk of alcohol use disorder under the two exposure conditions, averaged with respect to the covariate distribution.

For all estimators, SuperLearner was used for (initial) estimation of the conditional mean $\bar{Q}_0(A, W)$ and the propensity score $P_0(A=1|W)$. The library for $\bar{Q}_0(A, W)$ included the unadjusted mean, stepwise logistic regression, all logistic regressions with a single main term, and all logistic regressions with main term for the exposure and one additional covariate. The library for estimation of the propensity score $P_0(A=1|W)$ included the unadjusted mean, all logistic regressions with a single main term as well as stepwise logistic regressions with and without interactions. We chose these libraries for interpretability and to avoid overfitting. For rTMLE, the bounds on $\bar{Q}_0(A, W)$ were set to $[0, 8.5\%]$, as informed by a previous nationwide study on alcohol abuse and dependence (Hasin et al. 2007). Analyses based on upper bounds of 7.5% and 10% yielded nearly identical results. As before, inference was based on the Abadie-Imbens standard error estimator for the PSM estimator (Abadie and Imbens 2006; Sekhon 2011) and the estimated influence curve for the others. We assumed the standardized estimators followed the standard normal distribution.

The results are presented in Table 2. The point estimates from the PSM estimator (0.66 %) and IPTW (0.82 %) were positive, but their confidence intervals were wide and included the null. The point estimates from AIPTW (0.61 %), TMLE (0.88 %) and rTMLE (0.85 %) were also all positive, and the confidence intervals for the TMLEs did not include the null. While the double robust estimators are expected to perform similarly asymptotically, their finite sample performance is expected to differ. In particular, both TMLEs benefit from being substitution estimators (Stitelman and van der Laan 2010; Gruber and van der Laan 2010; Sekhon et al. 2011; Petersen et al. 2012; Gruber and van der Laan 2013; Lendle et al. 2013). rTMLE further benefits by using model knowledge on the bounds of the mean of the rare outcome. Overall, the results suggest that there is an increased risk of alcohol use disorder among neighborhoods with more permissive drunkenness norms. This finding is in line with previous work in the population, which suggested a significant association between neighborhood norms about drinking and binge drinking (Ahern et al. 2008).

5 Discussion

In this paper, we proposed a new TMLE for evaluating causal effects and estimating associations with very rare outcomes and high dimensional data. The rare outcomes TMLE (rTMLE) is based on harnessing knowledge in the semiparametric model $\mathcal{M}$, which bounds the conditional mean outcome $\bar{Q}_0(A, W)$ from below by $\ell = 0$ and from above by $u < 1$. These bounds can be based on subject matter knowledge or selected with cross-validation. Estimators, incorporating this model knowledge, are expected to be more robust and precise in finite samples.

In simulations, the proposed rare outcomes TMLE performed as well or outperformed the alternative estimators. The PSM estimator and IPTW were biased under misspecification of

the propensity score. The standard TMLE algorithm suffered from bias and poor confidence interval coverage when the adjustment set for the conditional mean outcome was large. Both AIPTW and rTMLE were robust to model misspecification and yielded consistent estimates if either the conditional mean outcome or the propensity score were consistently estimated. AIPTW, however, is not a substitution estimator and yielded negative (impossible) risk estimates. In contrast, the proposed TMLE respected the global knowledge in the statistical model. Our simulations further highlighted the potential for data-adaptive estimation to avoid parametric assumptions and to increase power.

We focused on situations where the conditional mean of a rare outcome was bounded from below by $\ell = 0$ and from above by $u < 1$. The proposed TMLE is equally applicable to situations where the outcome is very common and thereby the conditional mean outcome is bounded from below by $0 < \ell$ and from above by $u = 1$. Furthermore, the rTMLE algorithm could be applied when we have knowledge of both the lower ℓ and upper bound u and possibly when there are different bounds for different strata of the exposure A and covariates W . We expect that incorporating this knowledge during estimation will increase stability and power especially when the sample size is small and the dimension of the covariates is high. This is an area of future work.

The proposed TMLE is easily generalizable for estimation of other parameters, including the risk ratios, odds ratios and the impacts of longitudinal exposures. The TMLE is also applicable to other sampling designs. Specifically, case-control studies are commonly employed to increase robustness and efficiency of the analysis of rare events. There are several well established methods that correct for selection on the outcome (Anderson 1972; Prentice and Breslow 1978; King and Zeng 2001; Robins 1999; Mansson et al. 2007). For example, van der Laan (2008) presented a general mapping of loss functions, substitution estimators and estimating equations developed for prospective sampling (i. e. cohort sampling) into loss functions, substitution estimators and estimating equations for biased sampling (e. g. case-control sampling). The estimator's properties, such as double robustness and asymptotic efficiency, are preserved under the mapping. As noted by van der Laan (2008), however, the sample size needed to detect effects on the order of the outcome prevalence will be very large, unless the conditional probability of the outcome is bounded from above by some small constant u . The estimator, proposed in this article, satisfies this condition by construction and thereby has the potential to achieve higher power than its unconstrained counterpart in finite samples. In other words, the mapping provided in van der Laan (2008) will allow us to weight rTMLE appropriately for case-control sampling. We expect the resulting estimator to offer an improvement in terms of stability and power in finite samples.

Acknowledgments

Funding: University of California, Berkeley Committee on Research, National Institutes of Health, (Grant/Award Number: 'DP2 HD 080350', 'R01 AI074345', 'R01 DA 017642', 'R01 DA 022720').

Appendix

Table 3

For Simulation 1, the number of negative estimates for the marginal risk under no exposure from AIPW in 2000 simulated data sets.

	$n = 1000$	$n = 2500$
Both Correct	120	5
$P_0(A = 1 W)$ misspecified	260	11
$\tilde{Q}_0(A, W)$ misspecified	130	6

Appendix A: Variance of the efficient influence curve at $P_0 \in \mathcal{M}$

Suppose the outcome Y is binary, and consider the semiparametric statistical model $\mathcal{M}$, bounding the conditional probability of the outcome $E(Y|A, W) = \tilde{Q}(A, W)$ by some $[\underline{\ell}, u]$ with $0 < \underline{\ell} < u < 1$. For simplicity, let us assume the lower bound $\underline{\ell}$ is 0 and re-express the conditional mean function as

$$\bar{Q}(A, W) = u\tilde{Q}(A, W)$$

for some $\tilde{Q}(A, W) \in [0, 1]$. Let $g_0(A|W)$ denote the conditional distribution of the exposure, given the covariates $P_0(A|W)$. Let $\Psi(\bar{Q}_0) = E_0[\bar{Q}_0(1, W) - \bar{Q}_0(0, W)]$ and $\Psi(\tilde{Q}_0) = E_0[\tilde{Q}_0(1, W) - \tilde{Q}_0(0, W)]$. Then the variance of the efficient influence curve at the true but unknown distribution P_0 can be written as

$$\begin{aligned} & \text{Var}[D^*(\bar{Q}_0, g_0)] \\ &= E_0 \left[\left(\frac{I(A=1)}{g_0(1|W)} - \frac{I(A=0)}{g_0(0|W)} \right)^2 (Y - \bar{Q}_0(A, W))^2 \right] + E_0 \left[(\bar{Q}_0(1, W) - \bar{Q}_0(0, W) - \Psi(\bar{Q}_0))^2 \right] \\ &= E_0 \left[\left(\frac{I(A=1)}{g_0(1|W)^2} + \frac{I(A=0)}{g_0(0|W)^2} \right) E_0 \left[(Y - \bar{Q}_0(A, W))^2 | A, W \right] \right] + E_0 \left[(\bar{Q}_0(1, W) - \bar{Q}_0(0, W) - \Psi(\bar{Q}_0))^2 \right] \\ &= E_0 \left[\frac{\bar{Q}_0(1, W)(1 - \bar{Q}_0(1, W))}{g_0(1|W)} + \frac{\bar{Q}_0(0, W)(1 - \bar{Q}_0(0, W))}{g_0(0|W)} \right] + E_0 \left[(\bar{Q}_0(1, W) - \bar{Q}_0(0, W) - \Psi(\bar{Q}_0))^2 \right] \\ &= u E_0 \left[\frac{\tilde{Q}_0(1, W)(1 - u\tilde{Q}_0(1, W))}{g_0(1|W)} + \frac{\tilde{Q}_0(0, W)(1 - u\tilde{Q}_0(0, W))}{g_0(0|W)} \right] + u^2 E_0 \left[(\tilde{Q}_0(1, W) - \tilde{Q}_0(0, W) - \Psi(\tilde{Q}_0))^2 \right]. \end{aligned}$$

Therefore, the variance of the efficient influence curve at $P_0 \in \mathcal{M}$ has been expressed as the upper limit u times a bounded function of means. In other words, the variance of the efficient influence curve is dampened by a factor of u , and the corresponding information for learning the target parameter from sample size n is amplified by a factor of $1/u$ (van der Vaart 1998).

Now consider the outcome Y to be a proportion. Specifically, suppose the outcome Y is the average of k independent Bernoulli's with probability $\tilde{Q}_0(A, W) \in [0, u]$. Then we have that the conditional mean and variance of Y are

$$E_0(Y|A, W) = \bar{Q}_0(A, W)$$

$$\text{Var}_0(Y|A, W) = \bar{Q}_0(A, W)(1 - \bar{Q}_0(A, W))/k.$$

The variance of the efficient influence curve at the true distribution P_0 is then

$$\begin{aligned} \text{Var}[D^*(\bar{Q}_0, g_0)] &= \frac{u}{k} E_0 \left[\frac{\tilde{Q}_0(1, W)(1 - u\tilde{Q}_0(1, W))}{g_0(1|W)} + \frac{\tilde{Q}_0(0, W)(1 - u\tilde{Q}_0(0, W))}{g_0(0|W)} \right] \\ &\quad + u^2 E_0 [(\tilde{Q}_0(1, W) - \tilde{Q}_0(0, W) - \Psi(\tilde{Q}_0))^2]. \end{aligned}$$

More generally, we can consider the outcome Y to be a rare bounded continuous outcome (e. g. the average of dependent Bernoulli's). Then we have

$$E_0[(Y - \bar{Q}_0(A, W))^2|A, W] = E_0[Y^2 - 2Y\bar{Q}_0(A, W) + \bar{Q}_0(A, W)^2|A, W] = E_0[Y^2|A, W] - \bar{Q}_0(A, W)^2.$$

Since the outcome Y is small on average and Y^2 is even smaller, we have that $E_0(Y^2|A, W)$ behaves as u^2 .

Appendix B: Asymptotic variance at a misspecified limit $\bar{Q}(A, W)$

Suppose we have n independent, identically distributed (i.i.d.) observations of $O = (W, A, Y)$ drawn from some P_0 in the semiparametric statistical model $\mathcal{M}$. Further suppose, for discussion, that the exposure mechanism $g_0(A|W)$ is known and satisfies the positivity assumption. Again, we consider Y to be binary for simplicity. Now consider two TMLEs $\Psi_{1,n}(P_n)$ and $\Psi_{2,n}(P_n)$, which are functions of the empirical distribution P_n . Suppose the first constrains the estimated risks $\bar{Q}_n(A, W)$ to be 1, while the second constrains its estimated risks $\bar{Q}_n(A, W)$ to be u . In other words, $\Psi_{2,n}(P_n)$ ensures the estimated risks are within the model bounds. Since both TMLEs solve the efficient influence curve equation, they are double robust and will be consistent even if $\bar{Q}_n(A, W)$ converges to a misspecified limit. The second TMLE, however, will often be more efficient when $\bar{Q}_0(A, W)$ is inconsistently estimated.

Under regularity conditions, the asymptotic variance of the first estimator $\Psi_{1,n}(P_n)$ is given by the variance of the efficient influence curve at the misspecified limit $\bar{Q}(A, W)$ divided by n :

$$\begin{aligned} n\text{Var}[\Psi_{1,n}(P_n)] &= \text{Var}[D^*(\bar{Q}_0, g_0)] + E_0 \left[\frac{(1 - g_0(1|W))}{g_0(1|W)} (\bar{Q}_0(1, W) - \bar{Q}(1, W))^2 \right] \\ &\quad + E_0 \left[\frac{(1 - g_0(0|W))}{g_0(0|W)} (\bar{Q}_0(0, W) - \bar{Q}(0, W))^2 \right] \\ &\quad + 2E_0[(\bar{Q}_0(1, W) - \bar{Q}(1, W))(\bar{Q}_0(0, W) - \bar{Q}(0, W))] \end{aligned}$$

where $\text{Var}[D^*(\bar{Q}_0, g_0)]$ is the variance of the efficient influence curve at $P_0 \in \mathcal{M}$ as given above (Proof available upon request). The second and third terms, involving squared deviations between the true mean $\bar{Q}_0(A, W)$ and limit $\bar{Q}(A, W)$, are always positive. The last term, involving the product of these deviations, will be positive when both treatment-specific means are under-estimated or over-estimated. Thereby, when $\bar{Q}(A, W)$ approaches 1 for some treatment-covariate combinations, these terms will be positive and contribute substantially to the asymptotic variance of the first estimator.

The second TMLE bounds the estimated mean $\bar{Q}_n(A, W)$ to be u . If we replace the true $\bar{Q}_0(A, W)$ with $u\bar{Q}_0(A, W)$ for some $\bar{Q}_0(A, W) \in [0, 1]$ and the limit $\bar{Q}(A, W)$ with $u\bar{Q}(A, W)$ for some $\bar{Q}(A, W) \in [0, 1]$, we can rewrite the asymptotic variance of the second estimator $\Psi_{2,n}(P_n)$ as

$$\begin{aligned} n\text{Var}[\Psi_{2,n}(P_n)] = & \text{Var}[D^*(\bar{Q}_0, g_0)] + u^2 E_0 \left[\frac{(1 - g_0(1|W))}{g_0(1|W)} (\bar{Q}_0(1, W) - \bar{Q}(1, W))^2 \right] \\ & + u^2 E_0 \left[\frac{(1 - g_0(0|W))}{g_0(0|W)} (\bar{Q}_0(0, W) - \bar{Q}(0, W))^2 \right] \\ & + 2u E_0 [(\bar{Q}_0(1, W) - \bar{Q}(1, W))(\bar{Q}_0(0, W) - \bar{Q}(0, W))] \end{aligned}$$

Since u is small, the contribution from misspecification of $\bar{Q}_0(A, W)$ is diminished in the second estimator, which enforces the constraints in the statistical model $\mathcal{M}$. Thereby, this estimator will be closer to achieving the efficiency bound even if $\bar{Q}_n(A, W)$ converges to a misspecified limit. This provides an asymptotic motivation for constructing a new TMLE, which guarantees the predicted probabilities are within model bounds and does not rely on them being nicely bounded by chance. Finally, we note that if $\bar{Q}_n(A, W)$ converges to the true $\bar{Q}_0(A, W)$, then the two estimators $\Psi_{1,n}(P_n)$ and $\Psi_{2,n}(P_n)$ will be asymptotically equivalent. Both estimators will achieve the efficiency bound in that their asymptotic variance will be given by the variance of the efficient influence curve at P_0 divided by sample size. Their finite sample performance, however, is still expected to differ.

Appendix C: R Code

Full R code is available on the author's website: http://works.bepress.com/laura_balzer/25.

References

- Abadie A, Imbens G. Large sample properties of matching estimators for average treatment effects. *Econometrica*. 2006; 74:235–267.
- Abadie, A., Imbens, G. Matching on the estimated propensity score. Technical report, NBER Technical Working Paper. 2015. <http://www.nber.org/papers/w15301>
- Ahern J, Galea S, Hubbard A, Midanik L, Syme SL. “Culture of drinking” and individual problems with alcohol use. *American Journal of Epidemiology*. 2008; 167:1041–1049. [PubMed: 18310621]
- Anderson J. Separate sample logistic discrimination. *Biometrika*. 1972; 59:19–35.
- Beck N, King G, Zeng L. Improving quantitative studies of international conflict: A conjecture. *American Political Science Review*. 2000; 94:21–25.
- Bickel, P., Klaassen, C., Ritov, Y., Wellner, J. Efficient and Adaptive Estimation for Semiparametric Models. Baltimore: Johns Hopkins University Press; 1993.

- Birthplace in England Collaborative Group. Perinatal and maternal outcomes by planned place of birth for healthy women with low risk pregnancies: The birthplace in England national prospective cohort study. *British Medical Journal*. 2011; 343:d7400. [PubMed: 22117057]
- Braitman L, Rosenbaum P. Rare outcomes, common treatments: Analytic strategies using propensity scores. *Annals of internal medicine*. 2002; 137:693–696. [PubMed: 12379071]
- Cepeda M, Boston R, Farrar J, Strom B. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American Journal of Epidemiology*. 2003; 158:280–287. [PubMed: 12882951]
- Concato J, Feinstein A, Holford T. The risk of determining risk with multivariable models. *Annals of internal medicine*. 1993; 118:201–210. [PubMed: 8417638]
- 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 Article 26.
- Gruber S, van der Laan M. An application of targeted maximum likelihood estimation to the meta-analysis of safety data. *Biometrics*. 2013; 69:254–262. [PubMed: 23379761]
- Harrell, F, Jr. *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis*. Berlin, Heidelberg, New York: Springer; 2001.
- Harrell F Jr, Lee K, Mark D. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*. 1996; 15:361–387. [PubMed: 8668867]
- Hasin D, Stinson F, Ogburn E, Grant B. Prevalence, correlates, disability, and comorbidity of DSM-IV alcohol abuse and dependence in the United States. *Archives of general psychiatry*. 2007; 64:830–842. [PubMed: 17606817]
- Hernán M, Brumback B, Robins J. Marginal structural models to estimate the causal effect of Zidovudine on the survival of HIV-positive men. *Epidemiology*. 2000; 11:561–570. [PubMed: 10955409]
- Joffe M, Rosenbaum P. Invited commentary: Propensity scores. *American Journal of Epidemiology*. 1999; 150:327–333. [PubMed: 10453808]
- King G, Zeng L. Logistic regression in rare events data. *Political Analysis*. 2001; 9:137–163.
- Lendle S, Fireman B, van der Laan M. Targeted maximum likelihood estimation in safety analysis. *Journal of Clinical Epidemiology*. 2013; 66:S91–S98. [PubMed: 23849159]
- Mansson R, Joffe M, Sun W, Hennessy S. On the estimation and use of propensity scores in case-control and case-cohort studies. *American Journal of Epidemiology*. 2007; 166:332–339. [PubMed: 17504780]
- McCullagh P. Quasi-likelihood functions. *Annals of Applied Statistics*. 1983; 11:59–67.
- National Institute on Alcohol Abuse and Alcoholism. Alcohol use disorder: a comparison between DSM-IV and DSM-5. 2013. NIH Publication No. 13–7999. <http://pubs.niaaa.nih.gov/publications/dsmfactsheet/dsmfact.pdf>
- Neyman J, Dabrowska DM, Speed TP. Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes (in polish). *Statistical Science*. 1923; 5:465–480. (1990).
- Patorno E, Glynn R, Hernández-Díaz S, Liu J, Schneeweiss S. Studies with many covariates and few outcomes: Selecting covariates and implementing propensity-score-based confounding adjustments. *Epidemiology*. 2014; 26:268–278.
- Pearl, J. *Causality: Models, Reasoning and Inference*. 2nd. New York: Cambridge University Press; 2000.
- Peduzzi P, Concato J, Kemper E, Holford T, Feinstein A. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*. 1996; 49:1373–1379. [PubMed: 8970487]
- Petersen M, Porter K, Gruber S, Wang Y, van der Laan M. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*. 2012; 21:31–54. [PubMed: 21030422]
- Polley E, van der Laan M. SuperLearner: Super Learner Prediction. 2013 <http://CRAN.R-project.org/package=SuperLearner>, rpackage version 2.0–10.
- Prentice R, Breslow N. Retrospective studies and failure time models. *Biometrika*. 1978; 65:153–158.

- R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2014. <http://www.R-project.org>
- RARECARENet. Information network on rare cancers. 2014 URL <http://www.rarecarenet.eu/rarecarenet/>.
- 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.
- Robins J. [Choice as an alternative to control in observational studies]: Comment. *Statistical Science*. 1999; 14:281–293.
- Robins, J. 1999 Proceedings of the American Statistical Association. Alexandria, VA: American Statistical Association; 2000. Robust estimation in sequentially ignorable missing data and causal inference models; p. 6–10.
- Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983; 70:41–55.
- Rubin D. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*. 1974; 66:688–701.
- Rubin D. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*. 1980; 75:591–593.
- Rubin DB. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*. 1978; 6:34–58.
- Sekhon J. Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software*. 2011; 42:1–52.
- Sekhon, J., Gruber, S., Porter, K., van der Laan, M. Propensity-score-based estimators and c-TMLE. In: van der Laan, M., Rose, S., editors. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, Dordrecht, Heidelberg, London: Springer; 2011. p. 343–364.
- Stitelman O, van der Laan M. Collaborative targeted maximum likelihood for time-to-event data. *The International Journal of Biostatistics*. 2010; 6 Article 21.
- van der Laan M. Estimation based on case-control designs with known prevalence probability. *The International Journal of Biostatistics*. 2008; 4 Article 17.
- van der Laan M, Polley E, Hubbard A. Super learner. *Statistical Applications in Genetics and Molecular Biology*. 2007; 6 Article 25.
- van der Laan, M., Robins, J. *Unified Methods for Censored Longitudinal Data and Causality*. New York, Berlin, Heidelberg: Springer-Verlag; 2003.
- van der Laan, M., Rose, S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, Dordrecht, Heidelberg, London: Springer; 2011.
- van der Laan M, Rubin D. Targeted maximum likelihood learning. *The International Journal of Biostatistics*. 2006; 2 Article 11.
- van der Vaart, A. *Asymptotic Statistics*. New York: Cambridge University Press; 1998.
- Vittinghoff E, McCulloch C. Relaxing the rule of ten events per variable in logistic and cox regression. *American Journal of Epidemiology*. 2007; 165:710–718. [PubMed: 17182981]
- Wedderburn R. Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*. 1974; 61:439–447.
- World Health Organization. Geneva: 2013. Global tuberculosis report 2013.

Figure 1.

Boxplots of the point estimates from various algorithms for sample sizes $n = 1000$ (left) and $n = 2500$ (right) for Simulation 1. *CC* indicates both the outcome regression and the propensity score are correctly specified. *CM* indicates the outcome regression is correctly specified, but propensity score misspecified. *MC* indicates the outcome regression is misspecified, but the propensity score correctly specified. The dashed line indicates the true value $\psi_0 = 1.3\%$.

Figure 2.

Attained confidence interval coverage and power for the estimators in Simulation 1. The unfilled and filled points denote sample sizes of $n = 1000$ and $n = 2500$, respectively. Circles indicate both the outcome regression and the propensity score are correctly specified. Triangles indicate the outcome regression is correctly specified, but propensity score misspecified. Squares indicate the outcome regression is misspecified, and propensity score correctly specified. The dashed lines indicate 95% confidence interval coverage and 80% power, respectively.

Table 1

Estimator performance over 2,000 simulations of $n = 100$ clusters in Simulation 2. The rows indicate the estimator with “SL” denoting when SuperLearner was used for initial estimation of $\tilde{Q}_0(A, W)$.

	Treatment effect: $\Psi_0 = 0.30\%$				Null: $\Psi_0 = 0\%$			
	Ψ_n (%) ^a	MSE ^b	Cov. ^c	Power	Ψ_n (%)	MSE ^b	Cov. ^c	α
PSM	0.31	1.9E-6	0.99	0.40	0.01	2.3E-6	0.98	0.02
IPTW	0.31	1.3E-6	1.00	0.00	0.00	1.4E-6	1.00	0.00
AIPW	0.30	1.0E-6	1.00	0.34	0.00	1.1E-6	1.00	0.00
TMLE	0.30	1.0E-6	1.00	0.40	0.00	1.1E-6	1.00	0.00
rTMLE	0.30	1.0E-6	1.00	0.40	0.00	1.1E-6	1.00	0.00
rTMLE _{SL}	0.30	7.4E-7	0.94	0.94	0.00	7.8E-7	0.94	0.06

Notes:

^a Average value of the parameter estimate.
^b Mean squared error (bias-squared + variance).
^c Confidence interval coverage.

Table 2

Point estimates, variance estimates and confidence intervals in the applied data example. $\Psi_n(a)$ denotes the estimated marginal risk under exposure level a : $\Psi_n(a) = E_0[E_0(Y|A = a, W)]$. All measures are in %.

	$\Psi_n(1)$	$\Psi_n(0)$	Ψ_n	σ_n^2	95 % CI
PSM			0.66	0.0038	(−0.54, 1.87)
IPTW	2.06	1.24	0.82	0.0040	(−0.42, 2.07)
AIPW	2.13	1.52	0.61	0.0015	(−0.14, 1.36)
TMLE	2.18	1.29	0.88	0.0015	(0.12, 1.64)
rTMLE	2.18	1.33	0.85	0.0015	(0.10, 1.60)