The International Journal of Biostatistics

Volume 7, *Issue* 1 2011 *Article* 17

A Targeted Maximum Likelihood Estimator for Two-Stage Designs

Sherri Rose, *University of California, Berkeley*
Mark J. van der Laan, *University of California, Berkeley*

Recommended Citation:
Rose, Sherri and van der Laan, Mark J. (2011) "A Targeted Maximum Likelihood Estimator for Two-Stage Designs," *The International Journal of Biostatistics*: Vol. 7: Iss. 1, Article 17.
DOI: 10.2202/1557-4679.1217

A Targeted Maximum Likelihood Estimator for Two-Stage Designs

Sherri Rose and Mark J. van der Laan

Abstract

We consider two-stage sampling designs, including so-called nested case control studies, where one takes a random sample from a target population and completes measurements on each subject in the first stage. The second stage involves drawing a subsample from the original sample, collecting additional data on the subsample. This data structure can be viewed as a missing data structure on the full-data structure collected in the second-stage of the study. Methods for analyzing two-stage designs include parametric maximum likelihood estimation and estimating equation methodology. We propose an inverse probability of censoring weighted targeted maximum likelihood estimator (IPCW-TMLE) in two-stage sampling designs and present simulation studies featuring this estimator.

KEYWORDS: two-stage designs, targeted maximum likelihood estimators, nested case control studies, double robust estimation

Author Notes: The authors would like to thank the editor and referees for their helpful contributions. This research was funded by NIH grant R01 A1074345-01.

1 Introduction

We consider two-stage sampling designs where one takes a random sample from a target population and measures V on each subject in the first stage. The second stage involves drawing a subsample from the original sample, collecting additional data on the subsample. The decision regarding selection into the subsample can be influenced by V. This data structure can be viewed as a missing data structure on the full-data structure X collected in the second-stage of the study.

Specifically, the observed data structure on a randomly sampled subject can be represented as $O = (V, \Delta, \Delta X)$, where V is included in X, and Δ denotes the indicator of inclusion in the second-stage sample. The sample is then represented as n i.i.d. copies $O_1, \ldots, O_n$ of O. One particular type of two-stage sample is a so-called "nested case-control" sample where the outcome Y is included in V and subjects are sampled conditional on Y. We propose an inverse probability of censoring weighted targeted maximum likelihood estimator (IPCW-TMLE) for the estimation of target estimands, such as causal effects, in two-stage sampling designs.

A TMLE is a general procedure for estimation of a target parameter of the data-generating distribution in semiparametric models, and, in particular, can be used for any censored data structure. It is a two-step method where one first obtains an estimate of the data-generating distribution, and then in the second step updates the initial fit in a bias-reduction step targeted toward the parameter of interest, instead of the overall density. The TMLE unifies the locally efficient double robust properties of estimating function based methodology with the properties of maximum likelihood estimation. TMLEs are loss-based well-defined, efficient, unbiased substitution estimators of the target parameter of the data-generating distribution. In this paper, we present general IPCW-TMLEs, and then apply it to nested case-control samples in simulations.

2 Literature

Two-stage designs, including nested case-control studies, have been discussed and developed in previous literature, including Neyman (1938), Cochran (1963), Mantel (1973), Kupper et al. (1975), Liddell et al. (1977), Thomas (1977), and Breslow et al. (1983). Advantages can include reduction in costs associated with collecting data on the entire cohort and minimal losses in efficiency (Ernster, 1994; Rothman and Greenland, 1998; Essebag et al., 2003; Hak et al.,

2004; Vittinghoff and Bauer, 2006). Much of the literature focuses on logistic regression for effect estimation (Breslow and Cain, 1988; Flanders and Greenland, 1991; Ernster, 1994; Barlow et al., 1999; Szklo and Nieto, 1999). Robins et al. (1994) presented the missingness framework for two stage designs, and (double robust augmented) inverse probability of treatment-weighted estimators. We also refer to van der Laan and Robins (2003) which provides an in-depth study and overview of double robust estimation for missing data and causal inference data structures.

Wang et al. (2009). A recent paper by Wang et al. (2009) presents causal effect estimators using estimating equation methodology where the outcome Y, exposure A, and a subset S of covariates W are measured in the first stage (V includes Y, A, and S). They consider the same two-stage design, where one measures $V = (S, Y, A)$ on everyone in the sample, and $X = (S, Y, A, W)$ on the subjects in the validation sample defined by $\Delta = 1$, where the missingness mechanism is known. The Wang et al. article focuses on estimation of $EY(a)$ under the consistency assumption $Y = Y(A)$, the randomization assumption, A is independent of $Y(a)$, given (W, S), so that $EY(a) = E_{S,W}E_{X,0}(Y \mid A = a, S, W)$, and a parametric model for the treatment mechanism $\Pi(S, W) = P(A = 1 \mid S, W)$. Please see the Appendix for a discussion of the relationship between the estimators presented in Wang et al. (2009) and IPCW-TMLE.

TMLE. TMLE was first presented in van der Laan and Rubin (2006) and covered in detail in a forthcoming text (van der Laan and Rose, 2011). Methodology for other types of case-control studies, including independent case-control designs and individually matched case-control studies, were first presented in van der Laan (2008b) and Rose and van der Laan (2008, 2009). We make the following remark regarding the problem of estimation of an additive causal effect of a treatment A on outcome Y, controlling for confounders W. When Y is continuous, the TMLE based on a quasi-log-likelihood loss function with a logistic regression submodel is recommended (Gruber and van der Laan, 2010). This choice of submodel is more robust than a TMLE based on the squared error loss function with a linear regression, due to the linear regression fluctuations not respecting global constraints. The double robust parametric regression estimators presented by Scharfstein et al. are special cases of TMLEs, as discussed in Rosenblum and van der Laan (2010) (Scharfstein et al., 1999, p. 1141). The class of estimators given in Rosenblum and van der Laan (2010) are not identical to, but are asymptotically equivalent to a class of estimators (Tsiatis, 2006, Section 5.4, p. 132).

DOI: 10.2202/1557-4679.1217

3 IPCW-TMLE in Two-Stage Samples

Recall that we consider two-stage sampling designs where one takes a random sample from a target population, measures V on each subject in this first stage, and draws a subsample where one collects additional data. Inclusion in the subsample can be influenced by V. This data structure is a missing-data structure on the full-data structure X collected in the second-stage. The observed data structure is $O = (V, \Delta, \Delta X)$, where V is included in X, and Δ denotes the indicator of inclusion in the second-stage sample. The sample can then be represented as n i.i.d. copies $O_1, \ldots, O_n$ of O.

Let $P_{X,0}$ be the true probability distribution of X, and let $\mathcal{M}^F$ be a statistical model for $P_{X,0}$. Let $\Psi^F : \mathcal{M}^F \to \mathbb{R}^d$ be the target parameter of the full-data distribution, so that $\psi_0^F = \Psi^F(P_{X,0})$ is the parameter of the true probability distribution of X. We will denote the efficient influence curve of Ψ^F at a full-data distribution P_X with $D^F(P_X)$.

Let $g_{\Delta,0}(\delta \mid X) = P_{X,0}(\Delta = \delta \mid X)$ be the conditional probability distribution of Δ, given X. We assume the missing at random (MAR) assumption which states that $g_{\Delta,0}(\delta \mid X) = g_{\Delta,0}(\delta \mid V)$, i.e., Δ is independent of X, given V. For notational convenience, let $\Pi_0(V) \equiv g_{\Delta,0}(1 \mid V)$. This missingness mechanism might be known, a model might be available, or no further assumptions are made beyond MAR. Either way, the missingness mechanism can be estimated from the data (Δ_i, V_i), $i = 1, \ldots, n$, extracted from the observations O_i, $i = 1, \ldots, n$.

The statistical model $\mathcal{M}$ for the probability distribution P_0 of O is now defined in terms of the full-data statistical model and the model on the missingness mechanism. The efficient influence curve of $\Psi^F(P_{X,0})$ as an identifiable parameter of P_0 will be denoted with $D^*(P_0) = D^*(P_{X,0}, \Pi_0)$. We wish to estimate ψ_0^F based on a sample of n i.i.d. observations $O_1, \ldots, O_n$ from $P_0 \in \mathcal{M}$.

3.1 TMLE

The TMLE is a general procedure for estimation of a target parameter of the data-generating distribution in semiparametric models (van der Laan and Rubin, 2006). It marries the locally efficient double robust properties of estimating function based methodology and the properties of maximum likelihood estimation. TMLEs are loss-based well-defined, efficient, unbiased substitution estimators of the target parameter of the data-generating distribution.

Suppose that, given n i.i.d. observations $X_1, \ldots, X_n$, $P_{X,n}^*$ is a TMLE of $P_{X,0}$, and $\Psi^F(P_{X,n}^*)$ is the corresponding TMLE of ψ_0^F. Specifically, let $L^F(P_X)(X)$ be a full-data loss function (e.g., log-likelihood loss function) so

that

$$P_{X,0} = \arg \min_{P_X \in \mathcal{M}^F} E_0 L(P_X)(X).$$

Let $P^0_{X,n}$ be an initial estimator of $P_{X,0}$, possibly a L^F-loss based super learner (van der Laan et al., 2007). In addition, let $\{P_X(\epsilon) : \epsilon\}$ be a parametric working submodel of $\mathcal{M}^F$ through P_X at $\epsilon = 0$ so that its score at $\epsilon = 0$ equals, or spans, the full-data efficient influence curve:

$$\frac{d}{d\epsilon} L(P_X(\epsilon)(X)\Big|_{\epsilon=0} = D^F(P_X)(X), \text{ a.e.}$$

Such a TMLE $P^*_{X,n}$ is then defined as follows. For $k = 1, \ldots, K$, one computes the amount of fluctuation:

$$\epsilon^k_n = \arg \min_{\epsilon} P^F_n L^F(P^{k-1}_{X,n}(\epsilon)),$$

for $P^{k-1}_{X,n}$, and one sets $P^k_{X,n} = P^{k-1}_{X,n}(\epsilon^k_n)$. Here P^F_n is defined as the empirical distribution of the full-data $X_1, \ldots, X_n$, and, for a function f of X and probability distribution P, we used the notation $Pf \equiv \int f(x)dP(x)$ This updating process is iterated until convergence is achieved, i.e., K is chosen so that $\epsilon^K_n \approx 0$. The final update $P^K_{X,n}$ is denoted with $P^*_{X,n}$, and is called the TMLE of $P_{X,0}$. By the score condition on the working fluctuation model, it follows that

$$P_n D^F(P^*_{X,n}) = 0.$$

3.2 IPCW-TMLE

Given the TMLE developed for the full-data structure, we propose estimating ψ_0 based on $O_1, \ldots, O_n$ with an IPCW-TMLE. This IPCW-TMLE is simply defined by the above procedure with the addition of weights $\Delta_i/\Pi_n(V_i)$ for observations $i = 1, \ldots, n$, where $\Pi_n(V)$ is an estimator of $\Pi_0(V) \equiv g_{\Delta,0}(1 \mid V)$. Thus, this IPCW-TMLE involves the following steps:

IPCW initial estimator. Computing an initial IPCW-loss based estimator $P^0_{X,n}$ (e.g., using super learning, van der Laan et al., 2007) based on, for example, the IPCW-loss function

$$L(P_X)(O) \equiv \frac{\Delta}{\Pi_n(V)} L^F(P_X)(X).$$

Typically, this initial estimator is obtained by providing the initial estimator of $P_{X,0}$ in the full-data TMLE the IPCW weights.

IPCW-TMLE. For $k = 1, \ldots, K$, one computes the amount of fluctuation:

$$\begin{aligned} \epsilon_n^k &= \arg\min_\epsilon P_n L(P_{X,n}^{k-1}(\epsilon)) \\ &= \arg\min_\epsilon \frac{1}{n}\sum_{i=1}^n \frac{\Delta_i}{\Pi_n(V_i)} L^F(P_{X,n}^{k-1}(\epsilon))(X_i), \end{aligned}$$

for $P_{X,n}^{k-1}$, and one sets $P_{X,n}^k = P_{X,n}^{k-1}(\epsilon_n^k)$. This updating process is iterated until convergence is achieved, i.e., K is chosen so that $\epsilon_n^K \approx 0$. The final update is denoted with $P_{X,n}^*$, and is called the IPCW-TMLE of $P_{X,0}$.

Estimator of the target parameter. Finally, one evaluates the target parameter $\psi_n^* = \Psi^F(P_{X,n}^*)$. This is the TMLE of ψ_0^F.

As is apparent from the above definition of IPCW-TMLE, IPCW-TMLE is a targeted minimum loss based estimator (also TMLE), the generalization of TMLE (van der Laan, 2008a; van der Laan et al., 2009; van der Laan and Rose, 2011), but with a loss function defined as IPCW full-data loss function, and a parametric submodel $P_X(\epsilon)$ with score $(\Delta/\Pi_0(V))D^F(P_X)$ at $\epsilon = 0$.

Since it solves the IPCW full-data efficient influence curve equation, the IPCW-TMLE has an influence curve equal to $(\Delta/\Pi_0(V))D^F(P_X^1)$ if Π_0 is known, and P_X^1 denotes the limit of $P_{X,n}^*$ (see next section). Double robustness properties of the full-data efficient influence curve are immediately inherited by the IPCW-TMLE. If $\Pi_0(V)$ is consistently estimated with a maximum likelihood estimator, the influence curve of the IPCW-TMLE equals $(\Delta/\Pi_0(V))D^F(P_X^1)$ minus its projection on the tangent space of the model used for Π_0. As shown below, if we use a nonparametric maximum likelihood estimator for Π_0 and the full-data model is nonparametric, then the IPCW-TMLE solves the actual efficient influence curve equation, so that the IPCW-TMLE is efficient if $P_X^1 = P_{X,0}$. As with any asymptotically linear estimator, an estimate of the asymptotic variance $\sqrt{n}(\psi_n^* - \psi_0^F)$ is given by the empirical variance of the estimated influence curve.

3.2.1 IPCW Full-Data Efficient Influence Curve Equation

By the score condition on the working fluctuation model and $\epsilon_n^K = 0$, it follows that this IPCW-TMLE solves the ICPW full-data efficient influence curve equation:

$$0 = \frac{1}{n}\sum_{i=1}^n \frac{\Delta_i}{\Pi_n(V_i)} D^F(P_{X,n}^*)(X_i) = 0.$$

If the full-data TMLE is double robust or has other robustness properties, then these properties will be inherited by this IPCW-TMLE under the assumption that Π_n is a consistent estimator of Π_0. If V is discrete (with finite support), then we propose using a nonparametric estimator Π_n of Π_0.

In this case, we have the following important result. If the full-data model is nonparametric, V is discrete, and the missingness mechanism is estimated nonparametrically, then it follows that the IPCW-TMLE actually solves the true efficient influence curve equation. The latter implies that, under appropriate regularity conditions, and if P^*_{Xn} is consistent for $P_{X,0}$,the IPCW-TMLE will be an asymptotically efficient estimator of ψ_0.

Proof of Result. Consider the statistical model $\mathcal{M}$ for the observed missing data structure O implied by a nonparametric full-data model $\mathcal{M}^F$, the MAR assumption, possibly a model for the missingness mechanism Π_0, and V is discrete. Let $\Psi : \mathcal{M} \to \mathbb{R}$ be the statistical target parameter of interest defined by $\Psi(P_{P_X,\Pi}) = \Psi^F(P_X)$. The efficient influence curve of of Ψ at $P_0 = P_{P_{X0},\Pi_0}$ can be represented as

$$D^*(P_{X,0},\Pi_0)(O) = \frac{\Delta}{\Pi_0(V)} D^F(P_{X,0}) - \left\{ \frac{\Delta}{\Pi_0(V)} - 1 \right\} E_0(D^F(P_{X,0}) \mid \Delta = 1, V),$$

where $D^F(P_{X,0})$ is the efficient influence curve of the full-data parameter $\Psi^F : \mathcal{M}^F \to \mathbb{R}$.

The IPCW-TMLE $P^*_{X,n}$ solves $0 = P_n \Delta/\Pi_n D^F(P^*_{X,n})$ for any choice of estimator Π_n of Π_0. If Π_n is a nonparametric estimator of Π_0, then it follows that we also have

$$0 = P_n \left\{ \frac{\Delta}{\Pi_n(V)} - 1 \right\} E_n(D^F(P^*_{X,n}) \mid \Delta = 1, V),$$

for any choice of estimator of the regression $E_0(D^F(P^*_{X,n}) \mid \Delta = 1, V)$. As a consequence, it follows that for nonparametric estimators Π_n of Π_0, and IPCW-TMLE $P^*_{X,n}$, the IPCW-TMLE solves the efficient influence curve equation:

$$0 = P_n D^*(P^*_{X,n}, \Pi_n).$$

We also note that, if we fit Π_0 with a logistic regression, use it as an offset, and add a covariate $E_n(D^F(P^0_{X,n}) \mid \Delta = 1, V)/\Pi_n(V)$ to update this logistic regression fit of Π_0, iterate this updating process of the missingness mechanism until convergence, then the resulting fit Π^*_n will also solve:

$$0 = P_n \left\{ \frac{\Delta}{\Pi^*_n(V)} - 1 \right\} E_n(D^F(P^0_{X,n}) \mid \Delta = 1, V).$$

This follows from the well known fact that the score of a univariate linear logistic regression working model logit $\Pi(\delta) = \text{logit}\ \Pi + \delta C$ for the coefficient δ in front of the univariate covariate $C(V)$, equals $C(V)(\Delta - \Pi(\delta)(V))$. For such clever fits of the missingness mechanism we also have that $(\Pi_n^*, P_{X,n}^*)$ solves the efficient influence curve estimating equation:

$$0 = P_n \frac{\Delta}{\Pi_n^*(V)} D^F(P_{X,n}^*) - \left\{ \frac{\Delta}{\Pi_n^*(V)} - 1 \right\} E_n(D^F(P_{X,n}^0) \mid \Delta = 1, V),$$

so that double robustness and asymptotic efficiency can still be derived.

The latter type of IPCW-TMLE is slightly more complex than the regular IPCW-TMLE since it now also requires fitting the regression $E_n(D^F(P_{X,n}^*) \mid \Delta = 1, V)$. However, this represents a minor increase in complexity since it only involves running a mean regression of the outcome $D^F(P_{X,n}^*)(X_i)$ on V_i among the observations with $\Delta_i = 1$.

3.2.2 Risk Difference Example

In this section we demonstrate the IPCW-TMLE for the simple full-data structure $X = (W, A, Y)$, with covariate vector W, binary exposure (or treatment) A, and binary outcome Y. The observed data structure for a randomly sampled subject is $O = (V, \Delta, \Delta X)$, where $V = Y$. The target parameter of the full-data distribution of X is given by $\Psi^F(P_{X,0}) = E_{X,0}[E_{X,0}(Y \mid A = 1, W) - E_{X,0}(Y \mid A = 0, W)]$ and the full-data statistical model $\mathcal{M}^F$ is nonparametric. The full-data efficient influence curve $D^F(Q_0, g_0)$ at $P_{X,0}$ is given by

$$\begin{aligned} D^F(Q_0, g_0) &= \left(\frac{I(A=1)}{g_0(1 \mid W)} - \frac{I(A=0)}{g_0(0 \mid W)} \right) (Y - \bar{Q}_0(A, W)) \\ &\quad + \bar{Q}_0(1, W) - \bar{Q}_0(0, W) - \Psi^F(Q_0), \end{aligned}$$

where $Q_0 = (\bar{Q}_0, Q_{W,0})$, $Q_{W,0}$ is the true full-data marginal distribution of W, $\bar{Q}_0(A, W) = E_{X,0}(Y \mid A, W)$, and $g_0(a \mid W) = P_{X,0}(A = a \mid W)$. The first term will be denoted by D_Y^F and the second term by D_W^F, since these two terms represent components of the full-data efficient influence curve that are elements of the tangent space of the conditional distribution of Y, given A, W, and the marginal distribution of W, respectively. That is, D_Y^F is the component of the efficient influence curve that equals a score of a parametric fluctuation model of a conditional distribution of Y, given (A, W), and D_W^F is a score of a parametric fluctuation model of the marginal distribution of W. Note that

$D_Y^*(Q,g)$ equals a function $H^*(A,W)$ times the residual $(Y-\bar{Q}(A,W))$, where

$$H^*(A,W) = \left(\frac{I(A=1)}{g(1 \mid W)} - \frac{I(A=0)}{g(0 \mid W)}\right).$$

IPCW initial estimator. We can estimate the marginal distribution of $Q_{W,0}$ with IPCW-MLE

$$Q_{W,n}^0 = \arg\min_{Q_W} \sum_{i=1}^{n} L^F(Q_W)(W_i)\frac{\Delta_i}{\Pi_n(Y_i)},$$

where $L^F(Q_W) = -\log Q_W$ is the log-likelihood loss function for the marginal distribution of W. Note that $Q_{W,n}$ is a discrete distribution that puts mass $1/\{n\Pi_n(Y_i)\}$ on each observation W_i in the sample for which W_i is observed (i.e., $\Delta_i = 1$). Suppose that, based on a sample of n i.i.d. observations X_i, we estimated $\bar{Q}_0$ with loss-based learning using the log-likelihood loss function $L^F(\bar{Q})(X) = -\log \bar{Q}(A,W)^Y(1-\bar{Q}(A,W))^{1-Y}$. Given the actual observed data, we can estimate $\bar{Q}_0$ with super learning and weights $\Delta_i/\Pi_n(Y_i)$ for observations $i = 1,\ldots,n$, which corresponds to the same super learner but now based on the IPCW-loss function

$$L(\bar{Q})(O) \equiv \frac{\Delta}{\Pi_n(Y)} L^F(\bar{Q})(X).$$

Let $L^F(Q) = L^F(Q_W) + L^F(\bar{Q})$ be the full-data loss function for $Q = (\bar{Q}, Q_W)$ and let $L(Q,\Pi) = L^F(Q)\Delta/\Pi$ be the corresponding IPCW-loss function.

Similarly, we can estimate g_0 with loss-based super learning based on the IPCW-log-likelihood loss function

$$L(g)(O) \equiv \frac{\Delta}{\Pi_n(Y)}(-\log g(A \mid W)).$$

This now provides an initial estimator $Q_n^0 = (Q_{W,n}^0, \bar{Q}_n^0)$ and g_n^0. This estimator was obtained using the same algorithm for computing the initial estimator for the full-data TMLE, but now assigning weights $\Delta_i/\Pi_n(Y_i)$ to each observation. In essence, a full-data loss function $L^F(Q)$ for Q_0 used to obtain an initial estimator for the full-data TMLE has been replaced by the IPCW-loss function $L(Q,\Pi_n) = L^F(Q)\Delta/\Pi_n$, and, similarly, a full-data loss function $L^F(g) = -\log g$ has been replaced by $L(g,\Pi_n) = L^F(g)\Delta/\Pi_n$.

Parametric submodel for full-data TMLE. Let

$$Q^0_{W,n}(\epsilon_1) = (1 + \epsilon_1 D^F_W(Q^0_n))Q^0_{W,n}$$

be a parametric submodel through $Q^0_{W,n}$, and let

$$\bar{Q}^0_n(\epsilon_2)(Y = 1 \mid A, W) = \text{expit}\left(\log \frac{\bar{Q}^0_n}{(1 - \bar{Q}^0_n)}(A, W) + \epsilon_2 H^*_n(A, W)\right)$$

be a parametric submodel through the conditional distribution of Y, given A, W, implied by $\bar{Q}^0_n$. This describes a submodel $\{Q^0_n(\epsilon) : \epsilon\}$ through Q^0_n with a two-dimensional fluctuation parameter $\epsilon = (\epsilon_1, \epsilon_2)$. We have that $d/d\epsilon L^F(Q^0_n(\epsilon))$ at $\epsilon = 0$ yields the two scores $D^F_W(Q^0_n)$ and $D^F_Y(Q^0_n, g^0_n)$, and therefore spans the full-data efficient influence curve $D^F(Q^0_n, g^0_n)$, a requirement for the parametric submodel for the full-data TMLE. This parametric submodel and the loss function $L^F(Q)$ now defines the full-data TMLE and this same parametric submodel with the IPCW-loss function $L(Q, \Pi) = L^F(Q)\Delta/\Pi$ defines the IPCW-TMLE.

The IPCW-TMLE. Define

$$\epsilon_n = \arg\min_\epsilon P_n \frac{\Delta}{\Pi_n} L^F(Q^0_n(\epsilon)),$$

and let $Q^1_n = Q^0_n(\epsilon_n)$. Note $\epsilon_{1,n} = 0$ which shows that the IPCW empirical distribution of W is not updated. Note also that $\epsilon_{2,n}$ is obtained by performing an IPCW logistic regression of Y on $H^*_n(A, W)$ where $\bar{Q}^0_n(A, W)$ is used as an offset, and extracting the coefficient for $H^*_n(A, W)$. We then update $\bar{Q}^0_n$ with logit $\bar{Q}^1_n(A, W) = \text{logit } \bar{Q}^0_n(A, W) + \epsilon^1_n H^*_n(A, W)$. The updating process converges in one step in this example, so that the IPCW-TMLE is given by $Q^*_n = Q^1_n$.

Estimator of the target parameter. Lastly, one evaluates the target parameter $\psi^*_n = \Psi^F(Q^*_n)$, where $Q^*_n = (\bar{Q}^1_n, Q^0_{W,n})$, by plugging $\bar{Q}^1_n$ and $Q^0_{W,n}$ into our substitution estimator

$$\psi^*_n = \frac{1}{n}\sum_{i=1}^n \left\{ \frac{\Delta_i}{\Pi_n(Y_i)} \left(\bar{Q}^1_n(1, W_i) - \bar{Q}^1_n(0, W_i)\right)\right\}.$$

This is the IPCW-TMLE of ψ^F_0.

3.2.3 Right Censoring

Suppose our full-data structure is a right-censored data structure and we conduct a nested case-control study. For example, we have that X might be defined as $X = (W, A, \tilde{T}, \Xi, Y^*)$, where W are covariates, A is an exposure of interest, $\tilde{T} = \min(T, C)$, T is the time to the event, C denotes a censoring variable, $\Xi = I(\tilde{T} = T)$ is a failure indicator, and $Y^* = (\tilde{T} \leq t, \Xi = 1)$ is an indicator of having an observed failure by endpoint t. Our missing data structure is given by $O = (\Delta, \Delta X, \tilde{T}, \Xi, Y^*)$, where $\Delta = 1$ denotes membership in the nested case-control sample.

A special feature of this right censored data structure is that one will define a case based on a binary random variable Y^* that is not the outcome of interest. For example, Y^* could represent observed death by year 5, which would be denoted $Y^* = (\tilde{T} \leq 5 \text{ years}, \Xi = 1)$. It is important to stress that the definition of a case $(Y^* = 1)$ in a nested case-control study within a right censored data structure is therefore different than without right censoring. Let's say our parameter of interest $\Psi^F(P_{X,0})$ is the causal risk difference under causal assumptions: $E_{X,0}[P_{X,0}(T > 5 \mid A = 1, W) - P_{X,0}(T > 5 \mid A = 0, W)]$.

We define the TMLE for the full-data structure and we then use the IPCW-TMLE for actual missing data structure. In other words, we need a TMLE of ψ_0^F based on X, and then IPCW-TMLE is defined as well. The TMLE of the additive causal effect of treatment on survival, and other parameters, based on the right-censored data structure is presented elsewhere (Moore and van der Laan, 2009a,b; Stitelman and van der Laan, 2010; van der Laan and Rose, 2011).

3.2.4 Effect Modification

Nested case-control studies within clinical trials and observational studies are increasingly popular when researchers are interested in effect modification (Rothman and Greenland, 1998; Essebag et al., 2003, 2005; Prentice and Qi, 2006; Vittinghoff and Bauer, 2006; Polley and van der Laan, 2009). This is of particular importance when the candidate patient characteristic effect modifier of the treatment effect is difficult or expensive to measure (Vittinghoff and Bauer, 2006).

The general approach involves defining our full-data structure, for example, $X = (W, A^*, A, Y)$, and our observed data $O = (V, \Delta, \Delta X)$, where again V is in X. We are interested in studying the effect modification of a variable denoted A^*. Our full-data parameter of interest might be

$$\tilde{\psi}_0^F \quad = \quad E_0[\bar{Q}_0(1,1) - \bar{Q}_0(1,0) - \bar{Q}_0(0,1) + \bar{Q}_0(0,0)],$$

where $\bar{Q}_0(a^*, a) = E_0(Y \mid A^* = a^*, A = a, W)$. The full-data TMLE involves first running an initial regression of Y on A^*, A, and W. We note that A and A^* are implicitly assumed to have finite support. The targeting step requires a parametric working submodel to fluctuate the initial estimator and a choice of loss function. We use a clever covariate that will define this parametric working submodel. The clever covariate for $\bar{Q}_0^*(a^*, a)$ is given by

$$H^*(a^*, a) = \frac{I(A^* = a^*, A = a)}{g_0(a^*, a \mid W)},$$

where $g_0(a^*, a \mid W) = P_{X,0}(A = a \mid W) P_{X,0}(A^* = a^* \mid A = a, W)$, and $P_{X,0}(A = a \mid W)$ may be known, as in a clinical trial, but $P_{X,0}(A^* = a^* \mid A = a, W)$ must be fitted. The clever covariate for the difference parameter $\tilde{\psi}_0^F$ is the corresponding difference of clever covariates. As loss function one can use the least squares loss function, in which case the working submodel is a linear regression of Y on H^* using the initial estimator as offset. If Y is binary, or continuous in $(0, 1)$ (e.g., after a linear transformation), then one can use the more robust quasi-log-likelihood loss function (Gruber and van der Laan, 2010). In the latter case, the working submodel is a logistic linear regression of Y on H^*, using the initial estimator as offset. Therefore, one can target the parameter with a single clever covariate, or one can target all four parameters with a four dimensional clever covariate, and look at multiple differences. This now defines the full-data TMLE for the desired target parameter $\tilde{\psi}_0^F$. The desired IPCW-TMLE for the observed data is obtained by assigning weights $\Delta_i / \Pi_n(Y_i)$ to each observation, or equivalently, by replacing the full-data loss function in the full-data TMLE by the IPCW-loss function.

4 Simulations

We present several simulation studies to examine the performance of the IPCW-TMLE. First, we generate simulated nested case-control samples within real cohort data. We then study the IPCW-TMLE in simulated cohorts.

4.1 SPPARCS Simulations

The National Institute of Aging funded Study of Physical Performance and Age-Related Changes in Sonomans (SPPARCS) is a population-based, census-sampled, study of the epidemiology of aging and health. Participants of this longitudinal cohort were recruited if they were aged 54 years and over and were

Table 1: **SPPARCS variables**

Variable		Description
Y		Death occurring within 5 years of baseline
A		LTPA score $\geq$ 22.5 METs at baseline
	HEALTH.EX	Health self-rated as "excellent"
	HEALTH.FAIR	Health self-rated as "fair"
	HEALTH.POOR	Health self-rated as "poor"
	SMOKE.CURR	Current smoker
	SMOKE.EX	Former smoker
W	*CARDIAC*	Cardiac event prior to baseline
	CHRONIC	Chronic health condition at baseline
	AGE.1	$x \leq 60$ years old
	AGE.2	$60 < x \leq 70$ years old
	AGE.4	$80 < x \leq 90$ years old
	AGE.5	$x > 90$ years old
	FEMALE	Female

residents of Sonoma, CA or surrounding areas. Study recruitment of 2092 persons occurred between May 1993 and December 1994 and follow-up continued for approximately 10 years. One area of particular research interest for this data has been the effect of vigorous leisure-time physical activity (LTPA) on mortality in the elderly, which has been studied in a previous collaboration (Bembom and van der Laan, 2008) using marginal structural models. LTPA was calculated from answers to a detailed questionnaire where performed vigorous physical activities are assigned standardized intensity values in metabolic equivalents (METs). The recommended level of energy expenditure for the elderly is 22.5 METs.

The full-data structure is $X = (W, A, Y)$, where $Y = I(T \leq 5 \text{ years})$, T is time to the event death, A is a binary categorization of LTPA, and W are potential confounders. These variables are further defined in Table 1. The observed data structure on a randomly sampled subject can be represented as $O = (V, \Delta, \Delta X)$, where V is in X. Of note is the lack of any right censoring in this longitudinal cohort. The outcome (death within or at five years after baseline interview) and date of death was recorded for each subject. This information was available from a variety of sources, including death certificates. Our parameter of interest is the risk difference $\psi_0^F = P_{X,0}(Y_1 = 1) - P_{X,0}(Y_0 = 1)$, the average treatment effect of LTPA on mortality five years after baseline interview.

Table 2: **SPPARCS cohort results.** The TMLE was estimated in the SPPARCS cohort. Sample size was 2066, with 269 deaths five years from baseline interview and 1797 nondeaths. RD is risk difference, SE is standard error, and p is p-value

	Estimate	SE	p
RD	-0.054	0.012	< 0.001

The cohort was reduced to a size of $n = 2066$, as 26 subjects were missing LTPA values and/or self-rated health score (1.2% missing data). The prevalence of death was 13%, and the number of cases in the cohort sample was $nC = 269$. The TMLE was estimated on the full cohort sample, and the results are displayed in Table 2. Within TMLE, the machine learning Deletion/Substitution/Addition (DSA) algorithm was used to obtain an estimate of the functions $\bar{Q}_0 = P_{X,0}(Y = 1 \mid A, W)$ and $g_0 = P_{X,0}(A \mid W)$ since the functional form of the data was unknown. One could also use an ensemble approach, such as super learning (van der Laan et al., 2007). The estimated parameter of interest was highly significant, and indicates that physical activity at or above recommended levels decreases five-year mortality risk in this population by 5.4%.

Nested case-control simulations. We used this cohort study to simulate nested case-control study designs where an estimate of the missingness weights were obtained from the full cohort. Members of the nested case-control sample are denoted with $\Delta = 1$. Our observed data structure was defined as $O = (V, \Delta, \Delta X)$ and we had $V = Y$. Therefore, the missing data structure ignored those individuals with $\Delta = 0$, except for the purpose of estimating $\Pi_0(V)$.

Control individuals were randomly sampled from among those still alive five years from baseline interview, and assigned the value $\Delta = 1$. This was a simplified approach compared to an incidence-density design where individuals are sampled from those still at risk of death at the time a case becomes a case. Sampling was performed with various numbers of controls relative to the number of cases ($2nC$, $3nC$, and $4nC$). The empirical values for $P_{X,n}(\Delta = 1 \mid Y = 0)$, were 0.299, 0.446, and 0.608 for the three sample sizes. All cases (Y=1) were sampled with probability 1.

The cohort was resampled 1000 times. In each of the 1000 cohort resamples, one nested case-control study was extracted; those individuals with ($\Delta = 1$), allowing for ties (Bureau et al., 2008). The estimated values $\Pi_n(V)$ used in the weight vector were taken from their respective cohort resample. The IPCW-TMLE was estimated in each of the 1000 nested case-control samples, and the TMLE was estimated in the cohort samples. The DSA algorithm was

Table 3: **SPPARCS simulated nested case-control results.** IPCW-TMLEs were estimated in the nested case-control samples, and TMLEs were estimated in the cohort samples. RD is risk difference, SE is standard error, RE is relative efficiency compared to cohort RD, $nC = 269$ is number of cases, and nCo is number of controls

	Sample size	Estimate	RE
Cohort RD	2,066	-0.055	1.000
	$nCo = 2nC$	-0.101	0.319
Nested case-control RD	$nCo = 3nC$	-0.056	0.567
	$nCo = 4nC$	-0.051	0.789

used to obtain estimates of the functions $\bar{Q}_0$ and g_0. The relative efficiency of the nested case-control parameters are compared to the cohort parameter in Table 3, as well as average values for the parameter of interest. Relative efficiency of the nested case-control design improved as the number of controls increases. With an average of 4 controls per case (approximately 1076 of the 1797 available noncase subjects), the relative efficiency of the nested case-control design reached 78.9%.

4.2 Simulated Cohort

In the SPPARCS data simulations, we did not know the true value of the parameter of interest. It was important to have a completely objective way of defining the truth, and to then assess the performance of our estimator with respect to the truth. Therefore, we repeat the same simulation study, but now from a population we fully understand, as we know the value of the true ψ_0^F. The cohort was sampled from the target population of 1,000,000 individuals. We simulated a five-dimensional covariate $W = (W_j : j = 1, \ldots, 5)$, a binary exposure A, and indicator Y, where 1 indicated disease (or in the case of the SPPARCS data, death by 5 years from baseline interview). These variables were generated according to the following rules:

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

$$g_0(A \mid W) = \text{expit}(W_1 + W_2 + W_3 + W_4),$$

$$\bar{Q}_0(A, W) = \text{expit}(A - 4W_1 + AW_1 - 1.5W_2 + \sin(W_5)).$$

The true value for the risk difference was $RD = -0.061$ and the prevalence

Table 4: **Simulation data nested case-control results.** IPCW-TMLEs were estimated in the nested case-control samples and TMLEs were estimated in the cohort samples. RD is risk difference, SE is standard error, RE is relative efficiency compared to cohort RD, $nC = 296$ is number of cases, and nCo is number of controls

	Sample size	Estimate	RE
Cohort RD	2,066	-0.063	1.000
	$nCo = 2nC$	-0.045	0.411
Nested case-control RD	$nCo = 3nC$	-0.068	0.725
	$nCo = 4nC$	-0.069	0.788

The true value for the risk difference was $RD = -0.061$ and the prevalence of death was 13.3%. One cohort sample was taken with 2,066 individuals, and the estimated value of death prevalence was 14.3%. The number of cases in the cohort sample was $nC = 296$. Controls were randomly sampled from among the noncases in the original cohort at various sample sizes relative to the number of cases ($2nC$, $3nC$, and $4nC$), and assigned the value $\Delta = 1$. Noncases that were not sampled were assigned the value $\Delta = 0$. The values for $P_{X,n}(\Delta = 1 \mid Y = 0)$ were 0.330, 0.506, and 0.674 for the three sample sizes. All cases were assigned $\Delta = 1$.

Logistic regression was used to estimate the functions $\bar{Q}_0$ and g_0 since the functional form was known. The relative efficiency of the nested case-control parameters are compared to the cohort in Table 4, as well as average values for the parameter of interest. As before, relative efficiency of the nested case-control design improves as the number of controls increases. With an average of 4 controls per case, the nested design reaches a relative efficiency of 78.4%.

4.3 Simulated Clinical Trial

For a simulated clinical trial, 10,000 subjects were sampled and assigned a treatment A. The outcome of disease was assigned with $P_{X,0}(Y = 1 \mid W, A) = \text{expit}(3A - 4W_1 + W_3 - 12W_4 - 2W_5 + 2A\sin(W_3))$. Of the 10,000 subjects, 647 individuals developed disease (6.47%). The value of the effect modification parameter of interest in the full trial was $\tilde{\psi}_0^F = 0.016$. The full-data in the randomized controlled trial cohort was analyzed with a TMLE.

We proposed that the effect modifier of interest, $W_3 \equiv A^*$ was only measured in a nested case-control sample. Controls were randomly sampled from among the noncases in the original cohort at various sample sizes relative to the number of cases ($2nC$, $3nC$, $4nC$, and $5nC$), and assigned $\Delta = 1$. Noncases that were not sampled were assigned $\Delta = 0$. The values for $P_{X,n}(\Delta =$

Table 5: **Randomized controlled trial simulation data nested case-control results.** IPCW-TMLEs were estimated in the nested case-control samples and TMLEs were estimated in the full trial samples. SE is standard error, RE is relative efficiency compared to cohort RD, $nC = 647$ is number of cases, and nCo is number of controls

	Sample size	Estimate	RE
Full trial $\tilde{\psi}^F$	10,000	0.016	1.000
Nested case-control $\tilde{\psi}^F$	$nCo = 2nC$	0.024	0.142
	$nCo = 3nC$	0.022	0.253
	$nCo = 4nC$	0.019	0.517
	$nCo = 5nC$	0.016	0.864

$1 \mid Y = 0$) were 0.141, 0.210, 0.280, and 0.350 for the four sample sizes. All subjects with $Y = 1$ were assigned $\Delta = 1$.

An IPCW-TMLE was used to analyze the nested case-control samples. Multinomial regression was used with main terms to estimate the function $\bar{Q}_0$, representing a misspecified model. Due to the double robustness of the TMLE and IPCW-TMLE procedures, the estimates of the parameter of interest are consistent even when $\bar{Q}_0$ is misspecified. The values for $g_0(A^* \mid W)$ were known since it was a randomized controlled trial. Results are displayed in Table 5. The relative efficiency of the nested case-control design improves as the number of controls increases, and with 38.8% of the total trial participants we reach an efficiency of 86.4%.

5 Discussion

Two-stage sampling designs, including nested case-control sampling, are popular in many fields, including epidemiology. They have the potential to reduce the costs associated with collecting data on the full cohort with minimal losses in efficiency (Ernster, 1994; Rothman and Greenland, 1998; Hak et al., 2004; Vittinghoff and Bauer, 2006). We introduced the IPCW-TMLE for estimation of causal effects in two-stage sampling designs, with a focus on nested case-control sampling designs. In general, TMLE methodology can be used in conjunction with procedures that handle censoring, missingness, measurement error, and other persistent issues found in public health and medicine, in addition to adjusting for the missingness due to the two-stage sampling design.

Our simulated nested case-control studies within the SPPARCS data demonstrated 78.9% efficiency with an average of 4 controls per case. We had 78.4%

efficiency in our simulated nested case-control studies within a simulated cohort, again with an average of 4 controls per case. These results coincided with the conclusions of Ury (1975), which noted that as a general rule, 4 controls per case yields a relative efficiency of 80.0%. We also demonstrated the use of IPCW-TMLEs for nested case-control study designs within randomized controlled trials when interested in an effect modification research question. With less than 40% of the trial subjects, we reached an efficiency of 86.4% compared to the full trial.

Maintainers of large comprehensive databases that include adverse events often require researchers to pay for access, and cost almost always increases as the sample size requested increases. Thus, nested case-control studies are also a natural design for studies of safety with pharmaceutical drugs. The IPCW-TMLE is maximally efficient in these scenarios as no covariate information on the noncase-control observations is discarded. With the increase in popularity of nested case-control study designs in longitudinal cohorts and randomized controlled trials, the IPCW-TMLE procedure provides an additional tool to yield unique biological and public health discovery.

Appendix: Wang et al. and IPCW-TMLE

Recall the paper by Wang et al. (2009) discussed in Section 2 where they consider the same two-stage design as in this paper. Let's consider the model for the observed data $O = (V, \Delta, \Delta X)$ implied by a nonparametric full-data model for the distribution of X, and known $P_{X,0}(\Delta = 1 \mid V)$. In that case, the IPCW-TMLE we propose is locally efficient if $P_{X,0}(\Delta = 1 \mid V)$ is nonparametrically estimated or is estimated in a targeted way as specified in our article, and will be inefficient otherwise. If the full-data model is not nonparametric, then our proposed IPCW-TMLE will not be locally efficient, even if $P_{X,0}(\Delta = 1 \mid V)$ is estimated nonparametrically.

If $X = (S, Y, A, W)$, and one only assumes the consistency and randomzation assumption, then the statistical model for the distribution of X is indeed nonparametric. Thus, in that statistical model, the proposed IPCW-TMLE of $EY(a)$ will be efficient if S, Y, A are discrete and $P_{X,0}(\Delta = 1 \mid S, Y, A)$ is estimated nonparametrically or in targeted manner. However, as in Wang et al., if one also assumes a parametric model for the treatment mechanism, then the statistical model for the full-data is *not* nonparametric. As a consequence of this choice of full-data model, the efficient influence curve does not exist in closed form, and has smaller variance than the efficient influence curve for the nonparametric full-data model, (and there exists a whole class of double robust

influence curves/estimating functions), so that the Cramer-Rao lower bound in their more restricted model is smaller than the Cramer-Rao lower bound for the nonparametric full-data model our IPCW-TMLE aims to achieve. For such a nonparametric full-data model, their locally efficient estimator solves the actual efficient influence curve estimating equation while the IPCW-TMLE solves the inefficient IPCW-full-data efficient equation.

Wang et al. also consider the subclass of influence functions/estimating functions generated by the nonparametric full-data model corresponding with a saturated parametric model for the treatment mechanism, and they refer to the optimal influence function in this subclass as the efficient double robust estimating function. Their efficient double robust estimating function equals the efficient influence curve for the observed data model implied by nonparametric full-data model, i.e., the efficient influence curve of our model. As a consequence, their efficient double robust estimator (based on solving the efficient double robust estimating equation) and our double robust TMLE are both locally efficient for the observed data model corresponding with the nonparametric full-data model. If the full-data model is nonparametric, V is continuous, and we do not use the targeted estimator of the missingness mechanism then our proposed IPCW-TMLE is not locally efficient, while their efficient double robust estimator will be locally efficient.

References

W.E. Barlow, L. Ichikawa, D. Rosner, and S. Izumi. Analysis of case-cohort designs. *J Clin Epidemiol*, 52(12):1165–1172, 1999.

O. Bembom and M.J. van der Laan. Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators. *Technical Report 230, Division of Biostatistics, University of California, Berkeley*, 2008.

N.E. Breslow and K.C. Cain. Logistic regression for two-stage case-control data. *Biometrika*, 75(1):11–20, 1988.

N.E. Breslow, J.H. Lubin, and P. Marek. Multiplicative models and cohort analysis. *J Am Stat Assoc*, 78:1–12, 1983.

A. Bureau, M.S. Diallo, J.M. Ordovas, and L.A. Cupples. Estimating interaction between genetic and environmental risk factors: Efficiency of sampling designs within a cohort. *Epidemiology*, 19(1):83–93, 2008.

W.G. Cochran. *Sampling Techniques*. Wiley, New York, NY, 1963.

V.L. Ernster. Nested case-control studies. *Prev Med*, 23(5):587–590, 1994.

V. Essebag, J. Genest Jr., S. Suissa, and L. Pilote. The nested case-control study in cardiology. *American Heart Journal*, 146(4):581–590, 2003.

V. Essebag, R.W. Platt, M. Abrahamowicz, and L. Pilote. Comparison of nested case-control and survival analysis methodologies for analysis of time-dependent exposure. *BMC Medical Research Methodology*, 5(5), 2005.

W.D. Flanders and S. Greenland. Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine*, 10(5), 1991.

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

E. Hak, F. Wei, D.E. Grobbee, and K.L. Nichol. A nested case-control study of influenza vaccination was a cost-effective alternative to a full cohort analysis. *J Clin Epidemiol*, 57(9):875–880, 2004.

L.L. Kupper, A.J. McMichael, and R. Spirtas. A hybrid epidemiologic study design useful in estimating relative risk. *J Am Stat Assoc*, (70):524–528, 1975.

F.D.K. Liddell, J.C. McDonald, and D.C. Thomas. Methods of cohort analysis: appraisal by application to asbestos mining. *J R Stat Soc Ser A*, (140):469–491, 1977.

N. Mantel. Synthetic retrospective studies and related topics. *Biometrics*, 29 (3):479–486, 1973.

K. Moore and M.J. van der Laan. Application of time-to-event methods in the assessment of safety in clinical trials. In K.E. Peace, editor, *Design and Analysis of Clinical Trials with Time-to-Event Endpoints*. Chapman & Hall/CRC Biostatistics Series, 2009a.

K.L. Moore and M.J. van der Laan. Increasing power in randomized trials with right censored outcomes through covariate adjustment. *Journal of Biopharmaceutical Statistics*, 19(6):1099–1131, 2009b.

J. Neyman. Contribution to the theory of sampling human populations. *J. Am. Statist. Ass.*, 33:101–116, 1938.

E.C. Polley and M.J. van der Laan. Selecting optimal treatments based on predictive factors. In K.E. Peace, editor, *Design and Analysis of Clinical Trials with Time-to-Event Endpoints.* Chapman & Hall/CRC Biostatistics Series, 2009.

R.L. Prentice and L. Qi. Aspects of the design and analysis of high-dimensional snp studies for disease risk estimation. *Biostatistics*, 7(3):339–354, 2006.

J.M. Robins, A. Rotnitzky, and L.P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

S. Rose and M.J. van der Laan. Simple optimal weighting of cases and controls in case-control studies. *The International Journal of Biostatistics*, 4(1): Article 19, 2008.

S. Rose and M.J. van der Laan. Why match? Investigating matched case-control study designs with causal effect estimation. *The International Journal of Biostatistics*, 5(1):Article 1, 2009.

M. Rosenblum and M.J. van der Laan. Targeted maximum likelihood estimation of the parameter of a marginal structural model. *The International Journal of Biostatistics*, 6(2):2010.

K. Rothman and S. Greenland. *Modern Epidemiology.* Lippincott, Williams and Wilkins, Philadelphia, PA, 2nd edition, 1998.

D. O. Scharfstein, A. Rotnitzky, and J. M. Robins. Adjusting for non-ignorable drop-out using semiparametric nonresponse models, (with discussion and rejoinder). *Journal of the American Statistical Association*, 94:1096–1120 (1121–1146), 1999.

O.M. Stitelman and M.J. van der Laan. Collaborative targeted maximum likelihood for time-to-event data. *The International Journal of Biostatistics*, 6(1):Article 21, 2010.

M. Szklo and F.J. Nieto. *Epidemiology: Beyond the Basics.* Jones & Bartlett Publishers, Boston, MA, 2nd edition, 1999.

D.C. Thomas. Addendum to: "Methods of cohort analysis: appraisal by application to asbestos mining" by F.D.K. Liddell and J.C. McDonald and D.C. Thomas. *J R Stat Soc Ser A*, (140):469–491, 1977.

A.A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer, New York, 2006.

H.K. Ury. Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data. *Biometrics*, 31(3):643–649, 1975.

M.J. van der Laan. The construction and analysis of adaptive group sequential designs. Technical report 232, Division of Biostatistics, University of California, Berkeley, March 2008a.

M.J. van der Laan. Estimation based on case-control designs with known prevalence probability. *The International Journal of Biostatistics*, 4(1): Article 17, 2008b.

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

M.J. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York, 2011. `www.targetedlearningbook.com`.

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

M.J. van der Laan, E.C. Polley, and A.E. Hubbard. Super learner. *Technical Report 222, Division of Biostatistics, University of California, Berkeley*, 2007.

M.J. van der Laan, S. Rose, and S. Gruber. Readings on targeted maximum likelihood estimation. Technical report 254, Division of Biostatistics, University of California, Berkeley, 2009.

E. Vittinghoff and D.C. Bauer. Case-only analysis of treatment-covariate interactions in clinical trials. *Biometrics*, 62(3):769–776, 2006.

W. Wang, D. Scharfstein, Z. Tan, and E.J. MacKenzie. Causal inference in outcome-dependent two-phase sampling designs. *J.R. Statist. Soc. B*, 71(5): 947–969, 2009.