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

Year 2009 *Paper* 253

Causal Inference for Nested Case-Control Studies using Targeted Maximum Likelihood Estimation

Sherri Rose* Mark J. van der Laan†

*Division of Biostatistics, University of California, Berkeley, sherrirosephd@gmail.com

†University of California - Berkeley, laan@berkeley.edu

http://biostats.bepress.com/ucbbiostat/paper253

Causal Inference for Nested Case-Control Studies using Targeted Maximum Likelihood Estimation

Sherri Rose and Mark J. van der Laan

Abstract

A nested case-control study is conducted within a well-defined cohort arising out of a population of interest. This design is often used in epidemiology to reduce the costs associated with collecting data on the full cohort; however, the case control sample within the cohort is a biased sample. Methods for analyzing case-control studies have largely focused on logistic regression models that provide conditional and not marginal causal estimates of the odds ratio. We previously developed a Case-Control Weighted Targeted Maximum Likelihood Estimation (TMLE) procedure for case-control study designs, which relies on the prevalence probability q0. We propose the use of Case-Control Weighted TMLE in nested case-control samples, with either known q0 or q0 estimated from the full cohort. We show that this procedure is efficient for a reduced data structure, the data structure where covariate information is not collected or available on non-case-control subjects, and recognize that it is not fully efficient for the full data. However, in many common scenarios, the full data is not available, thus our procedure is maximally efficient for the data given. For statistical inference, we view the nested case-control sample as a missing data problem (Robins et al., 1994). Case-Control Weighted TMLE on the reduced data structure is illustrated in simulations for cohorts with and without right censoring and also effect modification in randomized controlled trials.

1 Introduction

Nested case-control studies are conducted within a well-defined cohort arising out of a population of interest. Typically, all of the subjects that develop disease in the cohort (i.e., the cases) are selected along with a random sampling of non-diseased subjects. Controls may be selected at the time each case becomes a case from the population without an event at that time but at risk for the event or at the end of the study. These two groups of subjects then comprise the nested case-control sample, where it is common for additional information to be collected, such as the exposure of interest (Mantel, 1973; Kupper et al., 1975; Liddell et al., 1977; Breslow et al., 1983; Rothman and Greenland, 1998). This design is increasingly used in public health, medicine, and genomics to study relationships between exposures and disease in large observational cohorts and effect modification in randomized controlled trials (Rothman and Greenland, 1998; Essebag et al., 2003, 2005). Nested designs may reduce the costs associated with collecting data on the full cohort with only a nominal loss in efficiency (Ernster, 1994; Rothman and Greenland, 1998; Hak et al., 2004; Vittinghoff and Bauer, 2006).

However, whether nested within a large observational cohort or a randomized controlled trial, the case-control study nested within the full cohort is biased since the proportion of cases in the sample is not the same as the population of interest. Methods for analyzing case-control studies have largely focused on logistic regression models (Breslow and Cain, 1988). These models provide conditional and not marginal (causal) estimates of the odds ratio. We have developed a Case-Control Weighted Targeted Maximum Likelihood Estimation (TMLE) procedure for case-control samples, which relies on the prevalence probability $q_0 \equiv P_0^*(Y = 1)$. TMLE is a general procedure for estimation, and can be used for any full data model and parameter of interest. It is a two-step method where one first obtains an estimate of the data-generating distribution and then in second stage updates the initial fit in a bias-reduction step targeted towards the parameter of interest, instead of the overall density. For case-control data, we simply employ the use of case-control weights in Case-Control Weighted TMLE. We propose the extension of Case-Control Weighted TMLE in nested case-control samples, with either known q_0 or q_0 estimated from the full cohort. We show that this procedure is efficient for a reduced data structure, the data structure where covariate information is not collected or available on non-case-control subjects, and recognize that it is not fully efficient for the full data. However, in many common scenarios, the full data is not available, thus our procedure is maximally efficient for the data given. For statistical inference, we view the nested case-control sample as a

missing data problem Robins et al. (1994). We are able to estimate a variety of parameters with Case-Control Weighted TMLE, including the marginal exposure effect adjusted for confounders. These parameters can be viewed as the analogues of causal inference parameters, but for observational data. We refer to these parameters as variable importance parameters if we are not willing to make causal assumptions. We illustrate Case-Control Weighted TMLE on the reduced data structure in simulations for cohorts with and without right censoring and also effect modification in randomized controlled trials.

2 Background

2.1 Literature and Existing Methodology

Nested case-control studies were introduced in Mantel (1973) and further discussed and developed in Kupper et al. (1975), Liddell et al. (1977), Thomas (1977), and Breslow et al. (1983). Advantages include reduction in costs associated with collecting data on the entire cohort, minimal losses in efficiency, and having the cases and controls come from the same population (Ernster, 1994; Rothman and Greenland, 1998; Essebag et al., 2003; Hak et al., 2004; Vittinghoff and Bauer, 2006). The latter is frequently not the case in independent case-control study designs. Nested case-control designs have also been shown to have similar estimates for parameters such as the standardized morbidity ratio when compared to an analysis of the full cohort (Liddell et al., 1977; Breslow et al., 1983; Lubin, 1986).

Much of the literature for analysis of nested case-control studies focuses on logistic regression models. The use of conditional logistic regression, treating the nested case-control study as a sample matched on time, is frequently discussed (Breslow and Cain, 1988; Flanders and Greenland, 1991; Ernster, 1994; Barlow et al., 1999; Szklo and Nieto, 1999). Samuelsen (1997) constructs pseudolikelihoods for nested case-control study designs using the conditional probability that a subject will be selected as a control to build a general parametric regression estimator and a semiparametric proportional-hazards estimator. Proportional hazards models have also been discussed elsewhere (e.g., Lubin, 1986). An important reference for our methodology is Robins et al. (1994). Their paper includes a discussion of a missingness framework for the estimation of inverse probability of treatment weighted (IPTW) marginal causal parameters for nested case-control study designs. We also refer to van der Laan and Robins (2003) which handles double robust estimation for missing data structures.

Beyond the types of parameters being estimated, the literature on the analysis of nested case-control study designs could further be divided loosely into three groups. One group analyzes the nested case-control sample as a case-control sample, ignoring the first stage of sampling the cohort, for example Barlow et al. (1999). The second group analyzes the nested case-control sample as a missing data structure, such as Robins et al. (1994). The third group straddles both of these groups, for example Breslow and Cain (1988). Our methodology falls within this third group. We estimate our parameter with information from only the case-control sample, but our inference respects the missing data structure. Our variance estimates incorporate both the variability due to sampling the cohort from the population of interest and the variability arising from drawing the case-control sample from the cohort.

An additional division in the literature could be drawn based on methods that rely on knowledge of the prevalence probability $q_0 \equiv P_0^*(Y = 1)$. For example, the methodology of Robins et al. (1994) requires only that q_0 be small. Our proposed methodology uses knowledge of q_0, or a reasonable estimate of q_0 approximated within the full cohort. The use of q_0 to eliminate the bias of case-control sampling designs has previously been discussed as update to a logistic regression model with the intercept $\log q_0/(1-q_0)$ (Anderson, 1972; Prentice and Breslow, 1978; Greenland, 1981; Morise et al., 1996; Wacholder, 1996). Adding the intercept $\log q_0/(1-q_0)$ yields the true logistic regression function $P_0^*(Y = 1 \mid A, W)$ (Anderson, 1972; Prentice and Pyke, 1979). A discussion of this updated logistic regression and its sensitivity to model misspecification can be found in Rose and van der Laan (2008). Similarly, there is a wealth of literature which discusses estimation in nested case-control studies with known sampling probabilities from the cohort, such as Borgan and Langholz (1993).

2.2 Case-Control Weighted TMLE

TMLE is a general methodology introduced in van der Laan and Rubin (2006). It is an efficient and double robust procedure that can estimate a variety of parameters of interest. We propose the use of Case-Control Weighted TMLE, which is simply a TMLE procedure that relies on the prevalence probability for case-control weights, in the case-control observations nested within a cohort. We will view the nested case-control sample within the cohort as a biased case-control sample in order to estimate our parameter of interest. Thus, here we discuss the general methodology for Case-Control Weighted TMLE before describing its application for use in nested case-control studies.

Case-Control Weighted TMLE, discussed in van der Laan (2008), maintains the locally efficient double robustness properties of estimating function based

methodology, and unifies maximum likelihood estimation (MLE) with estimating function methodology into a method improving on both. The case-control weighting framework maps estimation methods designed for non-case-control sampling into methods for case-control sampling. Case-control weighting allows us to provide TMLE methodology, which targets the parameter of interest, for biased case-control sampling in the form of Case-Control Weighted TMLE. Our procedure is a general methodology for the estimation of a parameter of a probability distribution, such as marginal causal effects and variable importance measures. The methodology relies on knowledge of the true prevalence probability $P_0^*(Y=1) \equiv q_0$, or a reasonable approximation, to eliminate the bias of the case-control sampling design.

Let us define $O^* \sim P_0^*$ as the experimental unit and corresponding distribution P_0^* of interest. To generalize, our case-control weighting maps a function of O^* into a function of the case-control data structure O, while preserving the expectation of the function. For example, the experimental unit of interest may be defined as $O^* = (W, A, Y) \sim P_0^*$, which consists of baseline covariates W, an exposure variable A, and a binary outcome Y. Then, in an independent case-control study design sampling can be described as first sampling (W_1, A_1) from the conditional distribution of (W, A), given $Y = 1$ for a case and then sampling J controls (W_0^j, A_0^j) from (W, A), given $Y = 0, j = 1, \ldots, J$. The observed data structure O is then defined by:

$$O = ((W_1, A_1), (W_0^j, A_0^j : j = 1, \ldots, J)) \sim P_0, \text{ with}$$

$$(W_1, A_1) \sim (W, A \mid Y = 1)$$
$$(W_0^j, A_0^j) \sim (W, A \mid Y = 0),$$

where the cluster containing one case and J controls is considered the experimental unit, and the marginal distribution of this cluster is specified by P_0^*. A case-control dataset of this design then consists of n i.i.d. observations $O_1, \ldots, O_n$ with sampling distribution P_0 as described above. The model $\mathcal{M}^*$, where q_0 may or may not be known, implies models for the marginal distribution of cases (W_1, A_1) and controls $(W_2^j, A_2^j), j = 1, \ldots, J$. Of note, if the independent case-control sampling design is conducted simply as sampling nC cases from the conditional distribution of (W, A), given $Y = 1$, and sampling nCo controls from (W, A), given $Y = 0$, the value of J used to weight each control is then nCo/nC.

Let $O^* \to D^*(O^*)$ represent an estimating function or loss function for O^* that can be used to estimate the parameter of interest of P_0^* based on an i.i.d. sample of O^*. We are concerned with mapping this function D^* into a function

for this same parameter of interest, but now based on sampling O (a biased sample for O^*). We define the case-control weighted version:

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

which is now a function of the observed experimental unit O. Additionally, we define the expectation operator $P_{0,q_0}D^* = P_0 D_{q_0}$, which takes the expectation of the case-control weighted function $D_{q_0}(O)$ with respect to P_0. Similarly, we define the empirical expectation $P_{n,q_0}D^* = P_n D_{q_0}$ as the empirical mean of the case-control weighted D_{q_0}, where P_n is the empirical distribution of $O_1, \ldots, O_n$. Now, we can let $D^*(O^*)$ be a function so that $P_0^* D^* \equiv E_{P_0^*} D^*(O^*) = 0$. Then $P_0 D_{q_0} = 0$, and

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

In more generality, for any function D^* and corresponding case-control weighted function D_{q_0}, we have

$$P_0 D_{q_0} = P_0^* D^*.$$

Given a model $\mathcal{M}^*$ for p_0^*, we can estimate P_0^* with a case-control weighted maximum likelihood estimator:

$$p_n^* = \arg\max_{p^* \in \mathcal{M}^*} \sum_{i=1}^{n} L(O_i, p^*),$$

where $L(O_i, p^*)$ is the case-control weighted log likelihood loss function for the density p_0^* of O^* under sampling of $O \sim P_0$:

$$L(O_i, p^*) = q_0 \log p^*(W_1, A_1, 1) + (1-q_0)\frac{1}{J}\sum_{j=1}^{J} \log p^*(W_2^j, A_2^j, 0).$$

Now, let $D^*(P_0^*)$ be the efficient influence curve of the parameter $\Psi^* : \mathcal{M}^* \to \mathbb{R}^d$. We consider an initial estimator P_n^{*0} of P_0^* based on $O_1, \ldots, O_n$ such as a case-control weighted maximum likelihood estimator according to a working model within $\mathcal{M}^*$. Let $\{P_n^*(\epsilon) : \epsilon\}$ be a submodel of $\mathcal{M}^*$ with parameter ϵ satisfying that the linear span of its score at $\epsilon = 0$ includes $D^*(P_n^{*0})$. Then we let ϵ_n^1 be the case-control weighted maximum likelihood estimator of ϵ:

$$\epsilon_n^1 = \arg\max P_{n,q_0} \log p_n^{*0}(\epsilon).$$

From this we obtain an update $P_n^{*1} = P_n^{*0}(\epsilon_n^1)$ of the initial estimator P_n^{*0}. This updating process is iterated until step k at which $\epsilon_n^k \approx 0$. The final update is denoted P_n^*. By the score condition, this final estimator solves the case-control weighted efficient influence curve:

$$0 = P_{n,q_0} D^*(P_n^*) = P_n D_{q_0}(P_n^*)$$

up to numerical precision (van der Laan and Rubin, 2006). We refer to $\psi_n = \Psi^*(P_n^*)$ as the case-control weighted targeted maximum likelihood estimator of ψ_0.

The theoretical development of Case-Control Weighted TMLE can be found in van der Laan (2008). In Rose and van der Laan (2008), we implemented Case-Control Weighted TMLE and presented a comparison of the procedure to an existing method for estimation of the causal parameters in case-control studies, the approximately correct IPTW of Robins (1999). We demonstrated that Case-Control Weighted TMLE outperforms the IPTW method for estimation of the marginal causal odds ratio in many practical situations.

3 Methodology for Nested Designs

Our goal is to apply Case-Control Weighted TMLE methodology to nested case-control designs. First, it is important to understand the statistical framework for the design. Nested case-control study designs have a missing data structure, as presented by Robins et al. (1994), and which we will discuss here. We will use a reduced data structure to estimate the parameter of interest with our proposed case-control weighted targeted maximum likelihood estimator. This estimator solves the efficient influence curve equation for the reduced data structure.

3.1 The Data Structure

Let O^* be a full data structure of the experimental unit O^* represents the data that ideally would be observed in order to answer the research question of interest. In most studies, however, one or more components of the full data are subject to one or more types of missingness, and only $O = \Phi(O^*, \delta)$ can be observed, where Φ is a known many-to-one mapping and δ denotes a missingness variable. Here, O^* represents data from the full cohort data and the missingness variable indicates membership in the nested case-control sample.

Suppose the full data structure is $O^* = (W, A, Y)$ with Y being a binary outcome of interest, A a binary exposure, and W a vector of covariates. Let us also suppose that the observed data structure for the nested case-control study is $O = (\delta, \delta O_1^*, O_2^*)$, where $O^* = (O_1^*, O_2^*)$. Particular examples are that $O_1^* = A$ and $O_2^* = (W, Y)$, or $O_1^* = (A, W)$, and $O_2^* = Y$. It is assumed that O_2^* always includes Y. The observations with $\delta = 1$ are the observations in the nested case-control sample within the cohort and have additional variables O_1^* measured. If $O_2^* = Y$, the missing data structure essentially ignores the non-case-control observations, except for the purpose of estimating $q_0 \equiv P_0^*(Y = 1)$. Covariate and exposure information is not available or is not measured. This case is particularly interesting since we can show that the case-control weighted targeted maximum likelihood estimator using only the case-control observations and the empirical estimate of q_0 obtained from the full cohort is a targeted maximum likelihood estimator for this particular missing data structure $(\delta, \delta(W, A), Y)$. If covariate information is measured and available for non-case-control subjects, this missing data structure ignores the information and therefore our estimator is not fully efficient.

We assume the coarsening at random (CAR) assumption: $\Pi(O^*) \equiv P_0^*(\delta = 1 \mid O^*) = P_0^*(\delta = 1 \mid O_2^*)$, and a special case is that $P_0^*(\delta = 1 \mid O_2^*) = P_0^*(\delta = 1 \mid Y)$ with $P_0^*(\delta = 1 \mid Y = 1) = 1$ and $P_0^*(\delta = 1 \mid Y = 0) = p$, where p is estimated empirically from the data. In this case the selection for the case-control sample is based upon the outcome Y. One might wish to choose p so that a single case $(Y = 1, \delta = 1)$ corresponds with J-controls $(Y = 0, \delta = 1)$, on average. If $q_0 = P_0^*(Y = 1)$, then $Jq_0 P_0^*(\delta = 1 \mid Y = 1) = (1 - q_0) P_0^*(\delta = 1 \mid Y = 0)$, which results in $p = \frac{Jq_0}{1-q_0}$.

3.2 Parameter of Interest

The statistical problem is then to estimate the parameter $\psi_0 = \Psi^*(P_0^*)$ of the population distribution $P_0^* \in \mathcal{M}^*$ of (W, A, Y), known to be an element of some specified model $\mathcal{M}^*$, based on the nested case-control data set $O_1, \ldots, O_n \sim P_0$. $O^* \sim P_0^*$, the experimental unit of interest, is not the observed experimental unit, due to the missing data structure. P_0^* now represents the full data distribution and P_0 is the distribution of the missing data structure with observed experimental unit $O = (\delta, \delta O_1^*, O_2^*) \sim P_0$. We focus on the case $O_2^* = Y$, where covariate information on non-case-control subjects is unavailable or ignored, and view this missing data structure as a biased case-control sampling design in order to estimate our parameter of interest. An example of a parameter of interest is the marginal exposure effect on the additive scale, which can also

be viewed as the causal risk difference:

$$\begin{aligned}\psi_{0,RD} &\equiv E_0^*\{E_0^*(Y \mid A=1,W) - E_0^*(Y \mid A=0,W)\} \\ &= E_0^*(Y_1) - E_0^*(Y_0) = P_0^*(Y_1=1) - P_0^*(Y_0=1).\end{aligned}$$

This definition requires the specification of the counterfactual outcomes Y_0 and Y_1 for binary A and $(W, A, Y = Y_A)$ as a time-ordered missing data structure on (W, Y_0, Y_1). For a causal interpretation, one must also make the randomization assumption: $\{A \perp Y_0, Y_1 \mid W\}$, meaning there are no unmeasured confounders. This parameter can also be viewed as a W-adjusted variable importance parameter, as previously mentioned, without the need to make causal assumptions. See van der Laan (2006) for this framework. We make use of the shorthand $Q_0^* = P_0^*(Y \mid A, W)$ and $g_0^* = P_0^*(A \mid W)$, the latter often referred to as the "treatment mechanism" but as the "exposure mechanism" in case-control studies.

3.3 The Estimator

TMLE is a general procedure for estimation, and can be used for any full data model and parameter of interest. It is a two-step method where one first obtains an estimate of the data-generating distribution and then in second stage updates the initial fit in a bias-reduction step targeted towards the parameter of interest, instead of the overall density. For case-control data we then simply add case-control weighting, using the prevalence probability. Here we will use Case-Control Weighted TMLE applied to nested case-control data using an estimate of q_0 from the full cohort. Again we focus on the case where O_2^* in the experimental unit $O = (\delta, \delta O_1^*, O_2^*) \sim P_0$ is equal to Y. We can show that the case-control weighted targeted maximum likelihood estimator using only the case-control observations and the empirical estimate of q_0 obtained from the full cohort is a targeted maximum likelihood estimator for this particular missing data structure $(\delta, \delta(W, A), Y)$. In this special case, the $D^*(Q, g, \Pi)$ we solve is the efficient influence curve (see Section 3.4). In other cases, for example when $O_2^* = (W, Y)$, we follow the same template for targeted maximum likelihood, where the case-control weighted log-likelihood is the criterion for fit.

Let us say we are still interested in the risk difference. We also let Q_n^0 be an initial estimator of $Q_0^* = P_0^*(Y \mid A, W)$, say the case-control weighted maximum likelihood estimator, or equivalently, the inverse probability of censoring weighted (IPCW) logistic regression estimator. In the IPCW estimator, the weights are $\frac{\delta}{\Pi}$. Now, we construct the ϵ-extension $\text{logit}Q_n^0(\epsilon) =$

$\text{logit}Q_n^0 + \epsilon h(g)(A,W)$, where $h(g)(A,W) \equiv \frac{A}{g_0^*(1|W)} - \frac{1-A}{g_0^*(0|W)}$, and we estimate ϵ with IPCW MLE. Alternatively, one puts the inverse probability of censoring weights in the ϵ-covariate: $\epsilon h(g)(A,W)\frac{\delta}{\Pi}$. Let $Q_n = Q_n^0(\epsilon_n)$. We now solve $\psi_{n,RD} = P_n \frac{\delta}{\Pi}(Q_{1n} - Q_{0n})$, where $Q_{1n} = Q_n(1,W)$ and $Q_{0n} = Q_n(0,W)$. Note that this corresponds with the case-control weighted empirical mean over W. So this estimator $\psi_{n,RD}$ corresponds exactly with the case-control weighted targeted maximum likelihood estimator proposed in van der Laan (2008), Rose and van der Laan (2008), and Rose and van der Laan (2009).

3.4 The Efficient Influence Curve

In order to estimate our parameter of interest, we view the missing data structure $(\delta, \delta(W,A), Y)$, where covariate information on subjects outside the nested case-control sample is unavailable or discarded, as a case-control sample. However, inference for this parameter must respect the missing data structure in order to account for the two sources of variability in the estimator. The first source of variance arises due to drawing the cohort from the target population, and the second source of variance arises from drawing the case-control sample from the cohort. If our inference treated the sample simply as a case-control sample, we would not be incorporating the additional variance arising from sampling the cohort from the population. Thus, for inference, we use an efficient influence curve that respects the missing data structure to obtain an estimate of the variance of our estimator. However, the efficient influence curve can also be used to construct closed form locally efficient double robust estimators by using it as an estimating function. The case-control weighted targeted maximum likelihood estimator discussed in the previous section solves the efficient influence curve equation for the missing data structure $(\delta, \delta(W,A), Y)$.

Our methodology for independent case-control study designs relies on knowledge of q_0, or a reasonable approximation of q_0, for appropriate statistical inference. In nested case-control samples we can easily estimate q_0 from the full cohort data. Inference for nested case-control study designs also requires the CAR assumption: $\Pi(O^*) \equiv P_0^*(\delta = 1 \mid O^*) = P_0^*(\delta = 1 \mid O_2^*)$. Let us say that we are still interested in the risk difference, but note that the derivation of the efficient influence curve and corresponding estimators generalizes to all other parameters of the full data distribution.

The efficient influence curve in the nonparametric full data model for $O^* = (W, A, Y)$ is given by:

$$D(O^*) = h(g)(A,W)(Y - Q(A,W)) + Q(1,W) - Q(0,W) - \Psi(Q),$$

where $h(g)(A,W) \equiv \frac{A}{g(1|W)} - \frac{(1-A)}{g(0|W)}$. We will represent $D = D_1 + D_2$, where

$D_1 = h(g)(Y - Q)$. The efficient influence curve for the missing data model is obtained through the following doubly robust IPCW mapping applied to the full data efficient influence curve D (see van der Laan and Robins (2003)):

$$D^* = \frac{\delta}{\Pi}\{D - E(D \mid \delta = 1, O_2^*)\} + E(D \mid \delta = 1, O_2^*).$$

This efficient influence curve can now be used to construct closed form locally efficient double robust estimators by using it as an estimating function. One will also be able to construct corresponding targeted maximum likelihood estimators. Here, we will focus on the $O_2^* = Y$-case. We have

$$\begin{aligned} D^*(Q, g, \Pi) &= \frac{\delta}{\Pi}\{h(Y - Q) + (Q_1 - Q_0)\} \\ &\quad -\frac{\delta}{\Pi}E(h(Y - Q) + Q_1 - Q_0 \mid \delta = 1, Y) \\ &\quad +E(h(Y - Q) + Q_1 - Q_0 \mid \delta = 1, Y) - \Psi(Q), \end{aligned}$$

where we use the notation $Q_1(W) = Q_0^*(1, W)$, $Q_0(W) = Q_0^*(0, W)$, and $Q = Q_0^*(A, W)$. This efficient influence curve can be decomposed as the sum of the following two components:

$$\begin{aligned} D_1^* &= \frac{\delta}{\Pi}\left(h(Y - Q) - E(h(Y - Q) \mid \delta = 1, Y)\right) + E(h(Y - Q) \mid \delta = 1, Y) \\ D_2^* &= \frac{\delta}{\Pi}\left(Q_1 - Q_0 - E(Q_1 - Q_0 \mid \delta = 1, Y)\right) + E(Q_1 - Q_0 \mid \delta = 1, Y) - \Psi(Q). \end{aligned}$$

We claim that D_1^* is a score of $dP(Y \mid A, W)$ and D_2^* is a score of $dP(W)$ in the observed likelihood factorization of $(\delta, \delta(W, A), Y)$, where the conditional expectation contributions, given $(\delta = 1, Y)$, are coming from the $dP(Y)$-factor.

Viewing $D^* = D^*(Q^*, g^*, \Pi, \psi)$ as an estimating function in ψ, setting $P_n D^*(Q_n, g_n, \Pi, \psi_n) = 0$ for given estimators Q_n, g_n of Q_0, g_0, yields the solution for the risk difference:

$$\begin{aligned} \psi_n &= P_n \frac{\delta}{\Pi}\{h(g_n)(Y - Q_n) + (Q_{1n} - Q_{0n})\} \\ &\quad -\left(\frac{\delta}{\Pi} - 1\right)\{E_n(h(g_n)(Y - Q_n) + Q_{1n} - Q_{0n} \mid \delta = 1, Y)\}. \end{aligned}$$

It is necessary for us to estimate the nuisance parameters:

$$\begin{aligned} E_n(h(Y - Q_n) \mid \delta = 1, Y = y) &= \frac{\sum_{i=1}^n I(\delta_i = 1, Y_i = y)h(A_i, W_i)(y - Q_n(W_i, A_i))}{\sum_{i=1}^n I(\delta_i = 1, Y_i = y)} \\ E_n(Q_{1n} - Q_{0n} \mid \delta = 1, Y = y) &= \frac{\sum_{i=1}^n I(\delta_i = 1, Y_i = y)(Q_{1n} - Q_{0n})(W_i)}{\sum_{i=1}^n I(\delta_i = 1, Y_i = y)}. \end{aligned}$$

Our case-control weighted targeted maximum likelihood estimator solves the IPCW weighted efficient influence curve equation:

$$0 = P_n \frac{\delta}{\Pi} \left\{ h(g_n)(Y - Q_n) + (Q_{1n} - Q_{0n}) - \Psi(Q_n) \right\}.$$

In our case-control study nested within the cohort sample, we estimate q_0 with $q_{0n} = \frac{1}{n} \sum_i I(Y_i = 1)$ and use the corresponding Π_n. Suppose we estimate $\Pi = P(\delta = 1 \mid Y = y)$ with the empirical proportion of δ among the observations with $Y_i = y$. Then:

$$0 = P_n \left(\frac{\delta}{\Pi_n} - 1 \right) \left\{ E_n(h(g_n)(Y - Q_n) + Q_{1n} - Q_{0n} - \psi_n \mid \delta = 1, Y) \right\}.$$

This follows by first conditioning on $Y = y$, and then noting that $P_n(\delta/\Pi_n(y) - 1 \mid Y = y) = 0$ for each $y \in \{0, 1\}$. By estimating Π with the empirical distribution of δ, it follows that this targeted maximum likelihood estimator ψ_n solves the efficient influence curve equation:

$$0 = P_n D^*(Q_n, g_n, \Pi_n, \psi_n).$$

Thus, our case-control weighted targeted maximum likelihood estimator, using the empirical proportions from the total cohort sample for q_0 and $1 - q_0$, actually solves this efficient influence curve equation for the missing data structure $(\delta, \delta(W, A), Y)$. In particular, we can use $D^*(Q^*, g^*, \Pi, \psi)$ as the influence curve under the assumption that g_0^* is correctly estimated. This influence curve can then be used calculate standard errors of the case-control weighted targeted maximum likelihood estimator. An estimate of the asymptotic variance of $\sqrt{n}(\psi_{n,RD} - \psi_0)$ using the efficient influence curve $D^*(Q^*, g^*, \Pi, \psi)$ is given by $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \hat{D}^{*2}$. A 95% Wald-type confidence interval for a parameter estimate $\hat{\psi}$ can be constructed as: $\hat{\psi} \pm z_{0.975} \frac{\hat{\sigma}}{\sqrt{n}}$ with a p-value calculated as $2[1 - \Phi(|\frac{\hat{\psi}}{\hat{\sigma}/\sqrt{n}}|)]$. Resampling based methods can also be implemented to estimate the standard error of the estimated parameter of interest.

We conclude that our proposed case-control weighted targeted maximum likelihood estimator with the empirical q_{0n} is a targeted maximum likelihood estimator for the missing data structure $(\delta, \delta(W, A), Y)$, and is thus a locally efficient procedure for that data. If in truth, as may often be the case, the non-case-control observations have covariate data, then one can use a more efficient double robust estimator using the above efficient influence curve and estimating the nuisance parameters.

4 Right Censoring

Let us say that our full data structure (the cohort) is a censored data structure. For example, O^* might be defined as $O^* = (W, A, \tilde{T}, \Delta, Y^*)$, where:

$$
\begin{gathered}
W \text{ are covariates,} \\
A \text{ is an exposure of interest,} \\
\tilde{T} = \min(T, C), \\
T \text{ is the time to the event } Y, \\
C \text{ denotes a censoring variable,} \\
\Delta = I(\tilde{T} = T), \text{ and} \\
Y^* = (\tilde{T} \leq t, \Delta = 1).
\end{gathered}
$$

We can apply our case-control weights to any data structure, and therefore O^* can be a censored data structure and we are still able to use our methods. Thus, suppose our observed data for this full data O^* is then $O = (\delta, \delta(W, A), \tilde{T}, \Delta, Y^*)$. Again, $\delta = 1$ denotes membership in the nested case-control sample. A special feature of this right censored data structure is that the true Y is not observed or a part of the full data. Instead, as noted, we have $Y^* = (\tilde{T} \leq t, \Delta = 1)$. For example, this could represent observed death by year 5, which would be denoted $Y^* = (\tilde{T} \leq 5 \text{ years}, \Delta = 1)$. The observed data structure for cases is then conditional on $(Y^* = 1)$. 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 very different than without right censoring, and accounting for this difference is not trivial. This distinction, and right censoring in general, is often overlooked in nested case-control study designs. The definition of q_0 is now $q_0 = P_0^*(\tilde{T} \leq t, \Delta = 1)$. Thus, by design we let $P_0^*(\delta = 1 \mid Y^* = 1) = 1$ and $P_0^*(\delta = 1 \mid Y^* = 0) = p$ and assume the CAR assumption $\Pi(O^*) \equiv P_0^*(\delta = 1 \mid O^*) = P_0^*(\delta = 1 \mid O_2^*)$.

Suppose we wish to compute a targeted maximum likelihood estimator for O^* of a parameter ψ_0, for example $\psi_0 = P_0^*(T_1 \leq 5 \text{ years}) - P_0^*(T_0 \leq 5 \text{ years})$, where $T_1 = (T \mid A = 1, W)$ and $T_0 = (T \mid A = 0, W)$. Thus we note that occurrence of disease conditioned upon in the case-control sampling does not need to be an outcome of interest. Targeted maximum likelihood estimators can handle both confounding as well as right censoring. To handle the right censoring, one might make use of censoring weights $\Delta / \bar{G}(\cdot)$, where $\bar{G}(\cdot)$ is the censoring mechanism, which can be estimated efficiently with a Kaplan-Meier curve (van der Laan and Rubin, 2007). Now suppose A is expensive to measure and can only be collected in a subsample of O^*. A nested case-control study might be performed. We can then implement a case-control

weighted targeted maximum likelihood estimator, as discussed in Section 3.3, with weights implied by $q_0 = P_0^*(\tilde{T} \leq 5 \text{ years}, \Delta = 1)$ in addition to the censoring weights. While simple to implement, this estimator is not a full TMLE due to the ad hoc IPCW weighting. Thus the case-control weighted IPCW TMLE is defined as the TMLE estimator for the full data structure weighting each observation $(W_i, A_i, \tilde{T}_i, \Delta_i, Y_i^*)$ with $\frac{\Delta_i q_0}{\bar{G}(\tilde{T}_i|A_i,W_i)}$ if $(Y_i^* = 1)$ and each of J corresponding control observations receive weight $\frac{\Delta_i(1-q_0)\frac{1}{J}}{\bar{G}(\tilde{T}_i|A_i,W_i)}$ if$(Y_i^* = 0)$.

An additional approach includes the use of the targeted maximum likelihood estimator presented in Moore and van der Laan (2009). This estimator involves first estimating a hazard of T given (A, W), expressing this hazard fit as a logistic regression or multiplicative intensity, and subsequently adding a time dependent covariate $h(t, A, W)$ as an epsilon extension. The epsilon coefficient in front of the clever covariate is fitted with standard logistic regression or Cox proportional hazards software, treating the initial hazard as an offset. This updating process of the conditional hazard is iterated until convergence. Once this updated hazard fit is determined with this iterative targeted maximum likelihood algorithm, one evaluates the conditional survival functions $S_{T|A=1,W}(5 \text{ years})$ and $S_{T|A=0,W}(5 \text{ years})$ and averages over W with respect to the empirical distribuiton of W. This is now the targeted maximum likelihood estimator of ψ_0, which needs to be case-control weighted by giving each observation with $(Y^* = 1)$ a weight q_0 and each control observation with $(Y^* = 0)$ a weight $(1 - q_0)\frac{1}{J}$. Note that this means each step in the above described TMLE algorithm, including the initial hazard estimation, needs to be case-control weighted.

5 Effect Modification

Nested case-control studies within clinical trials are becoming increasingly popular when researchers are interested in effect modification (Rothman and Greenland, 1998; Essebag et al., 2003, 2005; Polley and van der Laan, 2009). This is of particular importance when the patient characteristic that may modify the treatment effect is difficult or expensive to measure (Vittinghoff and Bauer, 2006). The Women's Health Initiative is an example of a well known study where the investigators' effect modification research question led to a nested case-control study design within a randomized controlled trial (Prentice and Qi, 2006). Researchers were interested in studying SNPs associated with coronary heart disease, stroke and breast cancer and hormone treatments in

their placebo controlled combined hormone trial cohort of over 16,000 women.

Suppose that within a randomized controlled trial we are interested in studying the effect modification of a particular patient characteristic, denoted W_i. The randomized controlled trial was designed with two treatment arms, $A \in \{0, 1\}$, where probability of assignment was $\pi = 0.5$. The disease outcome was binary $Y \in \{0, 1\}$ and the parameter:

$$\psi_0 \equiv E_0^*\{E_0^*(Y \mid A = 1, W) - E_0^*(Y \mid A = 0, W)\}$$

can be used to determine the average treatment effect. W indicates a multi-dimensional covariate $W = (W_i : i = 1, \ldots, m)$. However, our parameter of interest was an effect modification parameter. It represents the effect modification between $W_i \sim Bernoulli(\gamma = 0.5)$ and the treatment on the disease, while adjusting for the variables $W_{(-i)}$. This parameter of interest can be expressed:

$$\begin{aligned}\tilde{\psi}_0 &\equiv E_0^*\{[E_0^*(Y \mid A = 1, A^* = 1, W_{(-i)}) - E_0^*(Y \mid A = 0, A^* = 1, W_{(-i)})] \\ &- [E_0^*(Y \mid A = 1, A^* = 0, W_{(-i)}) - E_0^*(Y \mid A = 0, A^* = 0, W_{(-i)})]\},\end{aligned}$$

which can be written as:

$$\tilde{\psi}_0 \equiv E_0^*\{E_0^*(Z \mid A^* = 1, W_{(-i)}) - E_0^*(Z \mid A^* = 0, W_{(-i)})\}$$

since $\pi = 0.5$, where $Z = Y(A-(1-A))$, $A^* = W_i$, and $W_{(-i)}$ are the covariates that do not include W_i (van der Laan, 2006; Polley and van der Laan, 2009). The value of Z takes on three values, which follow a multinomial distribution:

$$Z = \begin{cases} +1 & \text{if } Y = 1 \text{ and } A = 1 \\ 0 & \text{if } Y = 0 \\ -1 & \text{if } Y = 1 \text{ and } A = 0. \end{cases}$$

The effect of A^* on Z, adjusted for all other covariates $W_{(-i)}$, the parameter $\tilde{\psi}_0$, can be estimated with targeted maximum likelihood estimation. This effect estimate can be considered a causal effect modifier, if one is willing to make the assumptions discussed in Section 3.2. Now suppose A^* can only be measured in stored blood products that were collected at the beginning of the trial, and the analysis of the stored blood products in the entire trial would be prohibitively expensive. A nested case-control design would then be a natural design to study the effect modification of A^* on Z. Suppose the full data structure was defined as $O^* = (W_{(-i)}, A^*, A, Y)$. Our observed missing data structure of the nested case-control sample would then be $O = (W_{(-i)}, \delta, \delta A^*, A, Y)$. An estimate of q_0 would come from the full data, the complete randomized controlled trial.

6 Safety Analysis

Maintainers of large comprehensive databases that include adverse events, such as the General Practice Research Database (GRPD) and The Health Improvement Network (THIN), often require researchers to pay for access to the data. Cost is based on a number of factors, but almost always increases as the number of subjects requested increases. Analysis of the entire cohort of data would be cost prohibitive. Thus, nested case-control studies are also a natural design for studies of safety with pharmaceutical drugs, and our case-control methodology has the potential to provide novel insight. Recent drug safety failures (e.g., Baycol, Vioxx, Ortho Evra, and Rezulin) have led to serious side effects and deaths in users. Additional post-market evaluation tools are necessary for detecting true adverse effects among the large number of reports of side effects and adverse outcomes stored in reporting databases, which are most commonly analyzed with logistic regression, producing only conditional estimates of the odds ratio (e.g., Yang et al. (2006)). In combination with the appropriate handling of multiple testing issues, Case-Control Weighted TMLE in nested case-control studies can play an important role in the detection of true adverse events. We highlight that these are scenarios where we only have data on the case-control observations. For example, if $O^* = (W, A, Y)$, then $O = (\delta, \delta(W, A), Y)$. Thus, our estimator is maximally efficient and very appropriate for these types of study designs since no covariate information (e.g. W) on the non-case-control observations is discarded.

7 SPPARCS Data Analysis & 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 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. The data structure $O^* = (W, A, Y)$, 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. Of note is the

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.

‡ LTPA is 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.

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 $\psi_0 = E_0^*(Y_1) - E_0^*(Y_0)$, the average treatment effect of LTPA on mortality five years after baseline interview.

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 estimated value for q_0 from the cohort was $q_{0n} = 0.130$, and the number of cases in the cohort sample was $nC = 269$. The variables used in our analysis are defined in Table 1. TMLE was performed 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 estimates of $Q_0^*(A, W)$ and $g_0^*(A \mid W)$ since the functional form of the data was unknown. Our estimated parameter of interest is highly significant, and indicates that physical activity at or above recommended levels decreases five-year mortality risk in this population by 5.4%. See Table 2.

7.1 SPPARCS Simulations

We used this longitudinal cohort study to simulate nested case-control study designs where an estimate of the prevalence probability for the weights is obtained from the full cohort. For example, let us say that our full data structure $O^* = (W, A, Y)$ and observed data $O = (\delta, \delta O_1^*, O_2^*)$, where $O^* = (O_1^*, O_2^*)$, are defined by the variables in Table 1. Since this nested case-control study is simulated inside a cohort with exposure and covariate information on all controls, let us also say we set $O_1^* = (A, W)$, and $O_2^* = Y$. The SPPARCS variables W, A, and Y continue to be defined by those described in Table 1. Members of the case-control sample are denoted with $\delta = 1$. The likelihood of a single observation is then written as:

$$dP_0^*(O) \quad = \quad \{dP_0^*(W)dP_0^*(A \mid W)dP_0^*(Y \mid A, W)\}^{\delta} dP_0^*(Y)^{1-\delta}.$$

Since $O_2^* = Y$, the missing data structure ignores those individuals with $\delta = 0$, except for the purpose of estimating $P_0^*(Y = 1)$.

7.1.1 Nested Case-Control Simulations

In order to form a control sample from the SPPARCS cohort for the nested case-control design, 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 at various sample sizes relative to the number of cases ($2nC$, $3nC$, and $4nC$). The empirical values for p in $\Pi(O^*) \equiv P_0^*(\delta = 1 \mid O^*) = P_0^*(\delta = 1 \mid O_2^*) = P_0^*(\delta = 1 \mid Y)$, with $P_0^*(\delta = 1 \mid Y = 1) = 1$ and $P_0^*(\delta = 1 \mid Y = 0) = p$, were 0.299, 0.446, and 0.608 for the three sample sizes. Non-cases that were not sampled were assigned the value $\delta = 0$. All cases were assigned $\delta = 1$.

The cohort was then 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. A simulation design such as this was also used in Bureau et al. (2008). The estimated values of q_0 for use in the case-control weights for the nested case-control samples were taken from their respective cohort resample. Case-Control Weighted TMLE was performed on each of the 1000 nested case-control samples and TMLE was performed on the cohort samples. The DSA algorithm was used to obtain estimates of $Q_0^*(A, W)$ and $g_0^*(A \mid W)$ since the functional form of the data was unknown. 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 improves as the number of controls increases. With an average of 4 controls per case (approximately 1076 of the 1797 available non-case subjects), the relative efficiency of the nested case-control design reached 78.9%.

7.1.2 Nested Case-Control Simulations with Right Censoring

For our simulations with right censored data, we generated an uninformative uniform censoring variable C, which led to 30.8% censored data in the full cohort data $O^* = (W, A, \tilde{T}, \Delta, Y^*)$. The definitions for $\tilde{T}, \Delta$, and Y^* are as described in Section 4, with W, A, and Y described in Table 1. The estimated value for q_0 from the cohort was $q_{0n} = 0.110$, and the number of cases in the cohort sample, defined by $Y^* = (\tilde{T} \leq 5 \text{ years}, \Delta = 1) = 1$, was $nC = 229$. Controls were sampled from the cohort from among those subjects who had $Y^* = 0$. The observed data for the nested case-control sample was defined as: $O = (\delta, \delta(W, A), \tilde{T}, \Delta, Y^*)$. Sampling was performed at various sample sizes relative to the number of cases as in the previous simulation, and the cohort was then 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. Values for p were 0.249, 0.371, and 0.494 for the three sample sizes. The cohort was analyzed with TMLE using IPCW weights defined as: $w_{IPCW} = \frac{I(C > \min(T,5))}{\bar{G}(\min(T,5))}$, where $\bar{G}(\cdot)$ is the censoring mechanism. The censoring mechanism can be estimated efficiently with a Kaplan-Meier curve (van der Laan and Rubin, 2007). The nested case-control samples were analyzed in a similar fashion, although we now also use IPCW weights and case-control weights in Case-Control Weighted TMLE. 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. Relative efficiency of the nested case-control design improves as the number of controls increases, although the nested case-control design does not reach the same high level of efficiency with 4 controls per case as our previous simulation without right censoring.

Table 2: **SPPARCS Cohort Results.** TMLE was performed on the SPPARCS cohort. Sample size was 2066, with 269 deaths five years from baseline interview and 1797 non-deaths. RD is Risk Difference, SE is Standard Error, and P is P-value.

	Estimate	SE	P
RD	-0.054	0.012	< 0.001

Table 3: **SPPARCS Simulated Nested Case-Control Results.** Case-Control Weighted TMLE was performed on the nested case-control samples, and TMLE was performed on 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
Case-Control RD	$nCo = 2nC$	-0.101	0.319
	$nCo = 3nC$	-0.056	0.567
	$nCo = 4nC$	-0.051	0.789

Table 4: **SPPARCS Simulated Nested Case-Control Results with Right Censoring.** Case-Control Weighted IPCW TMLE was performed on the nested case-control samples, and IPCW TMLE was performed on the cohort samples. RD is Risk Difference, SE is Standard Error, RE is Relative Efficiency Compared to Cohort RD, $nC = 229$ is number of cases, and nCo is number of controls.

	Sample Size	Estimate	RE
Cohort RD	2,066	-0.064	1.000
Case-Control RD	$nCo = 2nC$	-0.040	0.270
	$nCo = 3nC$	-0.040	0.310
	$nCo = 4nC$	-0.057	0.440

8 Additional Simulation Studies

8.1 Simulated Cohort

In the SPPARCS data simulations, we did not know the true value of the parameter of interest. It is therefore 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 exact same simulation study, but now from a population we fully understand, as we know the value of the true ψ. The cohort was sampled from the target population of 1,000,000 individuals. We simulated a 5-dimensional covariate $W = (W_i : i = 1, \ldots, 5)$, 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:

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

$$g_0^*(A \mid W) = \frac{1}{1+\exp(-(W_1+W_2+W_3+W_4))}$$

$$Q_0^*(A,W) = \frac{1}{1+\exp(-(-A-4W_1+AW_1-1.5W_2+\sin(W_5)))}.$$

The true value for the risk difference was $RD = -0.061$, and the true value for q_0 was $q_0 = 0.133$. One cohort sample was taken with 2,066 individuals, and the estimated value of q_0 taken from the cohort was $q_{0n} = 0.143$. The number of cases in the cohort sample was $nC = 296$. Controls were randomly sampled from among the non-cases in the original cohort at various sample sizes relative to the number of cases ($2nC$, $3nC$, and $4nC$), and assigned the value $\delta = 1$. Non-cases that were not sampled were assigned the value $\delta = 0$. The values for p were 0.330, 0.506, and 0.674 for the three sample sizes. All cases were assigned $\delta = 1$.

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. Weights for the case-control samples were taken from their respective cohort resample. Case-Control Weighted TMLE was performed on each of the 1000 nested case-control samples and TMLE was performed on the cohort samples. Logistic regression was used to estimate $Q_0^*(A,W)$ and $g_0^*(A \mid W)$ since the functional form was known. The relative efficiency of the nested case-control parameters are compared to the cohort in Table 5, 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%. Bias results can be seen in Figure 1.

8.2 Simulated Clinical Trial

As previously discussed, nested case-control studies within clinical trials are becoming increasingly common when researchers are interested in effect modification. Thus, we provide an additional illustrative example of our methods for this research question. The simulated target population contained 1,000,000 individuals with covariates W. For the clinical trial, 10,000 were sampled and assigned a treatment A. The outcome of disease was assigned with $Y = 1/(1+\exp(-(3A-4W_1+W_3-12W_4-2W_5+2A\sin(W_3))))$. 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 $\tilde{\psi}_0 = E_0^*\{E_0^*(Z \mid A^* = 1, W_{(-i)}) - E_0^*(Z \mid A^* = 0, W_{(-i)})\} = 0.016$. The full data in the randomized controlled trial cohort was analyzed with TMLE.

However, suppose that the effect modifier of interest, $W_3 \equiv A^*$, could only be measured in stored blood products, which is a very expensive process. Therefore, we could not measure $\tilde{\psi}_0$, as discussed in Section 5, in the entire trial and chose a nested case-control design. In order to simulate a nested case-control study within our simulated clinical trial data, controls were randomly sampled from among the non-cases in the original cohort at various sample sizes relative to the number of cases ($2nC$, $3nC$, $4nC$, and $5nC$), and assigned $\delta = 1$. Non-cases that were not sampled were assigned $\delta = 0$. The values for p were 0.141, 0.210, 0.280, and 0.350 for the four sample sizes. All subjects with $Y = 1$ were assigned $\delta = 1$. The resampling procedure was the same as our previous simulated designs. Case-Control Weighted TMLE was used to analyze the nested case-control samples. Multinomial regression was used with main terms to estimate $Q_0^*(A^*, W)$, and this represents a misspecified model. Due to the double robustness of the TMLE and Case-Control Weighted TMLE procedures, the estimates of the parameter of interest are consistent even when $Q_0^*(A^*, W)$ or $g_0^*(A^* \mid W)$ is misspecified. The values for $g_0^*(A^* \mid W)$ were known since it was a randomized controlled trial. Results are displayed in Table 6. 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%.

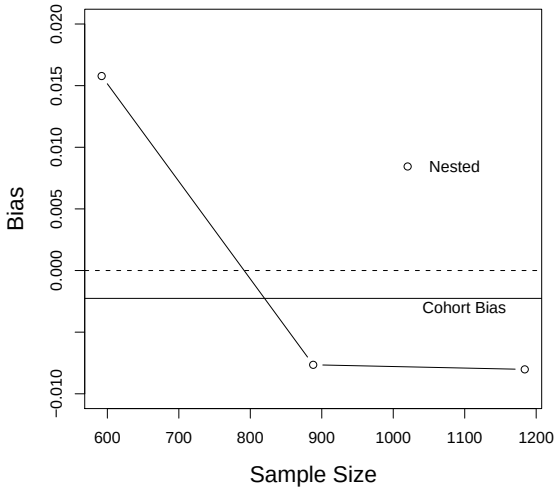

Figure 1: **Simulation Data Nested Case-Control – Bias Results for the Risk Difference.**

Table 5: **Simulation Data Nested Case-Control Results.** Case-Control Weighted TMLE was performed on the nested case-control samples and TMLE was performed on 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
Case-Control RD	$nCo = 3nC$	-0.068	0.725
	$nCo = 4nC$	-0.069	0.788

Table 6: **Randomized Controlled Trial Simulation Data Nested Case-Control Results.** Case-Control Weighted TMLE was performed on the nested case-control samples and TMLE was performed on 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 $\tilde{\psi}$	10,000	0.016	1.000
Case-Control $\tilde{\psi}$	$nCo = 2nC$	0.024	0.142
	$nCo = 3nC$	0.022	0.253
	$nCo = 4nC$	0.019	0.517
	$nCo = 5nC$	0.016	0.864

9 Discussion

Nested designs have the potential to significantly reduce the costs associated with collecting data on the full cohort with only minimal losses in efficiency (Ernster, 1994; Rothman and Greenland, 1998; Hak et al., 2004; Vittinghoff and Bauer, 2006). 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%. Our nested case-control simulations with right censoring within the SPPARCS data also demonstrated that methods for right censoring can be incorporated into the Case-Control Weighted TMLE procedure. In general, our case-control methodology can be used in conjunction with procedures that handle censoring, missingness, measurement error, and other persistent issues found in public health and medicine. We also demonstrated the use of Case-Control Weighted TMLE 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.

The extension of our Case-Control Weighted TMLE methodology to nested case-control study designs provides a double robust locally efficient estimation procedure for marginal causal effects and variable importance measures in nested designs. We showed that both the case-control weighted targeted maximum likelihood estimator and the IPCW estimator are targeted maximum

likelihood estimators for the missing data structure $(\delta, \delta(W,A), Y)$, and are thus locally efficient procedures for that data. For appropriate inference (e.g. construction of standard errors), however, the IPCW efficient influence curve must be implemented, or an appropriate resampling procedure such as bootstrapping. With the increase in popularity of nested case-control study designs in longitudinal cohorts and randomized controlled trials, the extension of our Case-Control Weighted TMLE procedure provides an additional tool to yield unique biological and public health discovery.

References

J.A. Anderson. Separate sample logistic discrimination. *Biometrika*, 59:19–35, 1972.

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.

O. Borgan and B. Langholz. Nonparametric estimation of relative mortality from nested case-control studies. *Biometrics*, pages 593–602, 1993.

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.

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. Greenland. Multivariate estimation of exposure-specific incidence from case-control studies. *J Chron Dis*, 34:445–453, 1981.

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.

J.H. Lubin. Extensions of analytic methods for nested and population-based incident case-control studies. *J Chronic Dis*, 39(5):379–388, 1986.

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, 2009.

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

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 N.E. Breslow. Retrospective studies and failure time models. *Biometrika*, 65(1):153–158, 1978.

R.L. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66:403–411, 1979.

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. [choice as an alternative to control in observational studies]: Comment. *Statistical Science*, 14(3):281–293, 1999.

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.

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

S.O. Samuelsen. A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika*, 84(2):379–394, 1997.

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.

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. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1):Article 2, 2006.

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

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 D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):Article 11, 2006.

M.J. van der Laan and D. Rubin. A note on targeted maximum likelihood and right censored data. *Technical Report 226, Division of Biostatistics, University of California, Berkeley*, 2007.

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

S. Wacholder. The case-control study as data missing by design: Estimating risk differences. *Epidemiology*, 7(2):144–150, 1996.

Y.X. Yang, J.D. Lewis, S. Epstein, and D.C. Metz. Long-term proton pump inhibitor therapy and risk of hip fracture. *JAMA*, 296(24):2947–2953, 2006.