The International Journal of Biostatistics

Volume 4, *Issue* 1 2008 *Article* 19

Simple Optimal Weighting of Cases and Controls in Case-Control Studies

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. (2008) "Simple Optimal Weighting of Cases and Controls in Case-Control Studies," *The International Journal of Biostatistics*: Vol. 4: Iss. 1, Article 19.
DOI: 10.2202/1557-4679.1115

Simple Optimal Weighting of Cases and Controls in Case-Control Studies

Sherri Rose and Mark J. van der Laan

Abstract

Researchers of uncommon diseases are often interested in assessing potential risk factors. Given the low incidence of disease, these studies are frequently case-control in design. Such a design allows a sufficient number of cases to be obtained without extensive sampling and can increase efficiency; however, these case-control samples are then 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 focused on utilizing logistic regression models that provide conditional and not causal estimates of the odds ratio. This article will demonstrate the use of the prevalence probability and case-control weighted targeted maximum likelihood estimation (MLE), as described by van der Laan (2008), in order to obtain causal estimates of the parameters of interest (risk difference, relative risk, and odds ratio). It is meant to be used as a guide for researchers, with step-by-step directions to implement this methodology. We will also present simulation studies that show the improved efficiency of the case-control weighted targeted MLE compared to other techniques.

KEYWORDS: case control sampling, causal effect, counterfactual, double robust estimation, estimating function, inverse probability of treatment weighting, locally efficient estimation, marginal structural models, targeted maximum likelihood estimation

Author Notes: The authors are most grateful to the reviewers for their insightful comments, which led to a major revision of this article. Both authors acknowledge support provided by NIH grant R01 A1074345-01.

1 Introduction

Case-control study designs are frequently used in public health and medical research to assess potential risk factors for disease. These study designs are particularly attractive to investigators researching rare diseases (i.e. probability of having the disease ≈ 0), as they are able to sample known cases of disease, versus following a large number of subjects and waiting for disease onset. Case-control studies can also yield increases in efficiency. However, case-control sampling is a biased sampling method; bias occurs due to the disproportionate number of cases in the sample versus the population of interest. Researchers commonly employ the use of a logistic regression model, treating the sample as a prospective sample, and estimate the *conditional* odds ratio of having disease given the exposure of interest and measured covariates. If one would like to estimate *marginal causal* effects, which correspond with the traditional parameters of interest in randomized trials, there is now a nonparametric double robust locally efficient procedure available. In van der Laan (2008), methodology for this marginal causal effect estimation theory in case-control designs is illustrated in detail. These techniques rely on knowledge of the true prevalence probability $P_0^*(Y = 1) \equiv q_0$ to eliminate the bias of the case-control sampling design. This methodology can be used in conjunction with other available procedures that handle censoring, missingness, measurement error, and other issues that are persistent in prospective and retrospective studies.

When possible, the population under study should be clearly defined. As such, the prevalence and incidence probabilities are then truly basic information about a population of interest. Given the availability of city, state, and national databases for many diseases, including many cancers, knowledge of the prevalence and incidence probabilities is now increasingly realistic. The literature, going back to the 1950's, supports this (see Cornfield (1951) and Cornfield (1956)). Nested case-control studies can also take advantage of the prevalence or incidence probability available in the full cohort study. The appropriateness of the use of the prevalence versus the incidence probability will depend on the nature of the case-control study design. The use of the these probabilities to eliminate the bias of case-control sampling design 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; Greenland, 2004). When the appropriate probability, or an estimate of the probability, is available, our procedure (van der Laan, 2008) can be implemented. In situations where data on the population of interest may be sparse, the use of a range for the probability is still beneficial.

An existing method for causal inference in case-control study designs, discussed by Robins (1999) and Mansson et al. (2007), involves the use of the exposure mechanism (also known as the propensity score or treatment mechanism in other literature) among control subjects as a weight to update a logistic regression of disease status on exposure. This inverse probability of treatment weighted (IPTW) marginal structural model does not require the knowledge of prevalence probability, only that the prevalence probability is close to zero. We will discuss this and other existing methods for analysis of case-control studies while stressing our new case-control weighting method that utilizes the prevalence probability.

The procedure, case-control weighted targeted maximum likelihood estimation, "targets" the parameter of interest rather than the distribution of interest. We use extra information, the estimate of the conditional distribution of the exposure given covariates among cases and controls (which we refer to as the exposure mechanism), to update an initial estimate of $P_0^*(Y \mid A, W)$. The procedure is double robust and locally efficient: it performs well as long as $P_0^*(Y \mid A, W)$ or $P_0^*(A \mid W)$ is correctly specified, is consistent if either of these models are correctly specified, and efficient if both are correctly specified. Our case-control weighting scheme successfully maps estimation methods designed for prospective sampling into methods for case-control sampling. It also produces efficient estimators when its prospective sample counterpart is efficient. For theoretical development of this new methodology, we will refer to van der Laan (2008). This article discusses case-control weighted targeted maximum likelihood for cumulative study designs with the prevalence probability and will focus on applications of the case-control weighting scheme in unmatched (independent) studies. For an extension of the methodology to matched case-control studies, see van der Laan (2008) and Rose and van der Laan (2008). Theory for incidence-density sampling with the incidence probability is also presented as an appendix in van der Laan (2008).

1.1 Case-Control Estimation

For ease of reference throughout the remainder of this article, we will present basic notation for understanding of the case-control estimation problem here. Let us define $O^* = (W, A, Y) \sim P_0^*$ as the experimental unit and corresponding distribution P_0^* of interest, which consists of baseline covariates W, an exposure variable A, and a binary outcome Y that defines case or control status. (For prospective studies, the exposure variable A would be referred to as the "treatment" variable.) P_0^* therefore represents the population from which all cases and controls will be sampled. One might be interested in several marginal causal effect parameters, including the causal risk difference, relative risk, and odds ratio. For causal effect parameter $\psi_0^* = \Psi^*(P_0^*) \in \mathbb{R}^d$ of $P_0^* \in \mathcal{M}^*$ and binary exposure $A \in \{0, 1\}$, these parameters are defined as:

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

$${}^{*}_{0,RR} = \frac{E_0^* E_0^*(Y \mid A = 1, W)}{E_0^* E_0^*(Y \mid A = 0, W)} = \frac{E_0^*(Y_1)}{E_0^*(Y_0)} = \frac{P_0^*(Y_1 = 1)}{P_0^*(Y_0 = 1)}, \tag{2}$$

and,

$${}^{*}_{0,OR} = \frac{P_0^*(Y_1 = 1)P_0^*(Y_0 = 0)}{P_0^*(Y_1 = 0)P_0^*(Y_0 = 1)}, \tag{3}$$

respectively. The causal versions of these definitions require 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), the full data structure. In addition, one must make the randomization assumption: $\{A \perp Y_0, Y_1 \mid W\}$. On the other hand, these parameters are always well defined parameters of the distribution of the data, and they can thereby be viewed as W-adjusted variable importance parameters without the need to make these assumptions. See van der Laan (2006) for this framework.

In van der Laan (2008), independent case-contol sampling is 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 $Y = 0, j = 1, \dots, J$. The observed data structure in independent case-control sampling is then defined by:

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

$$\begin{aligned} (W_1, A_1) &\sim (W, A \mid Y = 1) \\ (W_0^j, A_0^j) &\sim (W, A \mid Y = 0) \end{aligned}$$

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^*. Therefore, a case-control data set consists of n independent and identically distributed observations $O_1, \dots, O_n$ 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 $(W_2^j, A_2^j), j = 1, \ldots, J$.

This coupling formulation was useful when proving results for the case-control weighting methodology, and the tools provided in van der Laan (2008) show that the following is also true. If independent case-control sampling is described 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. This simple ratio $J = nCo/nC$ can be used effectively in practice.

2 Existing Methodology

As previously discussed, conditional estimation of the odds ratio of being diseased given the exposure of interest and baseline covariates is the prevalent method of analysis in case-control study designs. Key publications in the area of logistic regression for independent case-control study designs are Anderson (1972), Prentice and Pyke (1979), Breslow and Day (1980), and Breslow (1996). Greenland (1981) and Holland and Rubin (1988) discuss another model-based method: the use of log-linear models to estimate the marginal odds ratio. There are also several references for standardization in case-control studies, which estimates marginal effects with population or person-time averaging, including Rothman and Greenland (1998) and Greenland (2004). Benichou and Wacholder (1994) also present multivariate methods for population-based case-control studies. In this section, we will discuss the use of an intercept adjusted logistic regression as it can be incorporated into our case-control weighting framework. We will also discuss an IPTW marginal structural model for the estimation of causal effects as it is a related methodology, making use of the exposure mechanism. While these methods are discussed in current literature, they are infrequently implemented in current public health and medical research compared to the use of logistic regression for conditional effects.

2.1 Intercept Adjusted Logistic Regression

A thorough literature search yielded several publications suggesting the use of $\log q_0/(1 - q_0)$ as an update to the intercept of a logistic regression. (See Anderson (1972), Prentice and Breslow (1978), Greenland (1981), Morise et al. (1996), Wacholder (1996), and Greenland (2004), among others.) However, its use in practice remains limited. The adjustment is sometimes presented as a

DOI: 10.2202/1557-4679.1115

ratio of sampling fractions:

$$\log\left(\frac{P(\text{sampled} \mid Y=1)}{P(\text{sampled} \mid Y=0)}\right),$$

which reduces to $\log q_0/(1-q_0)$.

Adding the intercept $\log q_0/(1-q_0)$, denoted as $\log c_0$, yields the true logistic regression function $P_0^*(Y=1 \mid A, W)$ (Anderson, 1972; Prentice and Pyke, 1979). An intercept adjusted logistic regression can be used within the case-control weighting framework as an initial estimate of $P_0^*(Y \mid A, W)$. This will be discussed further in Section 3.2.1 and Section 4. The true logistic regression function can also be mapped to causal effect parameters by averaging over the case-control weighted distribution of W, which will also be discussed in Section 3.2.1.

2.2 IPTW

Robins (1999) and Mansson et al. (2007) discuss, under a rare disease assumption, the use of an approximately correct IPTW method in a marginal structural logistic regression model for case-control study designs. This procedure uses the estimated propensity score (exposure mechanism) among control subjects to update a logistic regression of Y on A. However, this IPTW estimator targets a nonparametrically non-identifiable parameter, which indicates strong sensitivity towards model misspecification for the exposure mechanism. See van der Laan (2008) for formal discussion of this result. Additionally, the causal effect estimates of the risk difference and relative risk cannot be obtained using this method. We also refer to Newman (2006) for a related IPTW-type approach for fitting marginal structural models based on case-control data. This method builds on the standardization approach in order to weight exposed and unexposed controls using a regression of A on W. We will include the IPTW method of Robins (1999) and Mansson et al. (2007) in our simulations.

3 Case-Control Weighted Targeted Maximum Likelihood Estimation

In this section, we provide the end user with a practical overview of the case-control weighting scheme for targeted maximum likelihood estimation in case-control study designs. For the formal statistical theory behind this technique, see van der Laan (2008). We discuss the implementation of case-control weighting for targeted maximum likelihood estimation both broadly and step-wise so that this article may be used as a guide to researchers wishing to employ these methods in their work.

3.1 Summary

Case-control weighted targeted maximum likelihood estimation for case-control study designs differs from other approaches to causal parameter estimation in case-control study design as it incorporates estimates of $P_0^*(Y \mid A, W)$, $P_0^*(A \mid W)$, and knowledge of q_0. Intercept adjusted logistic regression mapped to causal parameters discussed in the previous section relies on knowledge of only $P_0^*(Y \mid A, W)$ and q_0; the IPTW procedure of Robins (1999) and Mansson et al. (2007) relies on $P_0^*(A \mid W)$. The case-control weighted targeted maximum likelihood estimation procedure provides a nonparametric double robust locally efficient estimator: it performs well as long as $P_0^*(Y \mid A, W)$ or $P_0^*(A \mid W)$ is correctly specified, is consistent if either of these models are correctly specified, and efficient if both are correctly specified. It uses extra information, the estimate of the conditional distribution of the exposure given covariates among cases and controls, to update an initial estimate of $P_0^*(Y \mid A, W)$. One can use data-adaptive model-selection for estimation of $P_0^*(Y \mid A, W)$ and $P_0^*(A \mid W)$ within our procedure. (This will be discussed further in Section 5.) The procedure follows the basic steps enumerated below, which we then illustrate in more detail.

1. Assign weights q_0 to the cases and $(1 - q_0)\frac{1}{J}$ to the corresponding J controls.

2. Estimate the conditional probability of Y given A and W using assigned weights. The estimate of $P_0^*(Y \mid A, W) \equiv Q_0^*(A, W)$ is $\hat{Q}^*(A, W)$.

3. Estimate the conditional distribution of the exposure given covariates using assigned weights. The estimate of $P_0^*(A \mid W) \equiv g_0^*(A \mid W)$ is $\hat{g}^*(A \mid W)$.

4. Calculate the "clever covariate" for each subject based on $g_0^*(A \mid W)$. The covariate is estimated by $h(A, W)$.

5. Update the initial fit $\hat{Q}^*(A, W)$ from step 2 using the covariate $h(A, W)$. This is achieved by holding the coefficients of $\hat{Q}^*(A, W)$ fixed while estimating a new coefficient ϵ for $h(A, W)$ using weighted maximum likelihood estimation. The updated regression is given by $\hat{Q}_1^*(A, W)$

6. Use the assigned weights and $\hat{Q}_1^*(A,W)$ to estimate causal parameters of interest seen in formulas (1), (2) and (3). This is done by averaging over the case-control weighted distribution of W.

7. Calculate standard errors, and then, subsequently, p-values and confidence intervals, using the influence curve.

3.2 Implementation

The implementation of case-control weighted targeted maximum likelihood can be achieved using existing tools available in current software packages. Here we illustrate the steps described in Section 3.1.

3.2.1 Estimating $Q_0^*(A,W)$

After assigning weights q_0 and $(1-q_0)\frac{1}{J}$ to cases and controls, respectively, the first step in case-control weighted targeted maximum likelihood estimation for case-control designs is obtaining an estimate for $P_0^*(Y \mid A,W) \equiv Q_0^*(A,W)$. We offer two approaches for fitting this initial regression, the previously discussed intercept adjusted logistic regression, and a case-control weighted logistic regression. A comparison of these two approaches will be discussed in Section 4.

Intercept Adjusted Logistic Regression for $Q_0^*(A,W)$. Updating a logistic regression with $\log c_0$ is discussed in Section 2.1.

Case-Control Weighted Logistic Regression for $Q_0^*(A,W)$. Using the assigned weights, one simply performs maximum likelihood estimation for prospective sampling ignoring the case-control sampling design. If one considers a nonparametric model for the marginal distribution of the covariates and a model $\{Q_\theta^* : \theta\}$ for $Q_0^*(A,W)$, the case-control weighted maximum likelihood estimator for $Q_0^*(A,W)$ is then given by:

$$\hat{\theta} = \arg\max_\theta \sum_{i=1}^{n} q_0 \log \hat{Q}_\theta^*(A_{1i},W_{1i}) + (1-q_0)\frac{1}{J}\sum_{j=1}^{J} \log(1-\hat{Q}_\theta^*(A_{2i}^j,W_{2i}^j)).$$

Implementing case-control weighted maximum likelihood estimation, which is simply a weighted logistic regression, is quite straightforward, and can be done in many existing statistical software programs, including `SAS`, `STATA`, and `R`.

Outside of the case-control weighted targeted maximum likelihood estimation framework, case-control weighted logistic regression mapped to causal inference parameters produce efficient estimators. This mapping is accomplished by evaluating $\hat{Q}^*(A, W)$ at $A = 1$ and $A = 0$, applying the appropriate weights to estimate $P_0^*(Y_1 = 1)$ and $P_0^*(Y_0 = 1)$, and then computing the desired causal parameters of interest defined in formulas (1), (2), and (3). Estimating causal parameters will be discussed in more detail in Section 3.2.5. Case-control weighted logistic regression therefore provides researchers an immediate one-step intuitive procedure to estimate causal inference parameters in case-control study designs.

3.2.2 Estimating $g_0^*(A \mid W)$

The case-control targeted maximum likelihood estimation procedure uses the estimate of $Q_0^*(A, W)$ obtained above in conjunction with an estimate of $g_0^*(A \mid W)$. If one further considers a model $\{g_\eta^* : \eta\}$ for $g_0^*(A \mid W)$, the case-control weighted maximum likelihood estimator for $g_0^*(A \mid W)$ is given by:

$$\hat{\eta} = \arg\max_{\eta} \sum_{i=1}^{n} q_0 \log \hat{g}_\eta^*(A_{1i} \mid W_{1i}) + (1 - q_0)\frac{1}{J}\sum_{j=1}^{J} \log \hat{g}_\eta^*(A_{2i}^j \mid W_{2i}^j),$$

For improved performance of the targeted maximum likelihood estimator in a practical environment, estimated probabilites that are smaller than 0.01 can be set to 0.01 (Bembom et al., 2007).

3.2.3 Calculating $h(A, W)$

After estimating $Q_0^*(A, W)$ and $g_0^*(A \mid W)$, the next step requires calculation of a "clever covariate" for each subject. This covariate, which is calculated as if one has a prospective sample, takes the form:

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

for the risk difference. It is easy to see that for $A = 1$ the second term disappears, and for for $A = 0$ the first term disappears. Two covariates:

$$h_0(A, W) \equiv \left(- \frac{I(A = 0)}{\hat{g}^*(A = 0 \mid W)}\right) \text{ and } h_1(A, W) \equiv \left(\frac{I(A = 1)}{\hat{g}^*(A = 1 \mid W)}\right)$$

are used for estimation of other parameters, such as the relative risk and odds ratio. For a more detailed discussion of the "clever covariate," see van der Laan and Rubin (2006) and Moore and van der Laan (2007).

3.2.4 Updating $\hat{Q}^*(A,W)$

Updating $\hat{Q}^*(A,W)$ involves performing an additional weighted regression with $h(A,W)$ as a supplementary covariate. All other coefficients in the initial regression $\hat{Q}^*(A,W)$ are held fixed, and an intercept is suppressed in order to estimate the coefficient in front of $h(A,W)$, denoted ϵ. The case-control weighted targeted maximum likelihood estimation procedure is then able to incorporate information from $\hat{g}^*(A \mid W)$, through $h(A,W)$, into an updated regression. It does this by extracting $\hat{\epsilon}^1$, the case-control weighted maximum likelihood estimator of ϵ, from the fit defined above, and updating the regression estimate $\hat{Q}^*(A,W)$. This updated regression is then given by $\hat{Q}^*_1(A,W)$:

$$\hat{Q}^*_1(A,W) = \hat{Q}^*(A,W) + \hat{\epsilon}^1 h(A,W).$$

The updating procedure is iterated until convergence, although in many examples convergence is achieved in one step.

3.2.5 Estimating Causal Parameters

The risk difference, relative risk, and odds ratio, were previously defined generally in formulas (1), (2), and (3). The estimate $\hat{Q}^*_1(A,W)$ obtained in the previous step can be easily mapped into causal parameters of interest in the case-control weighting scheme for targeted maximum likelihood estimation by averaging over the case-control weighted distribution of W. This is accomplished by evaluating $\hat{Q}^*_1(A,W)$ at $A=1$ and $A=0$ and applying weights q_0 for cases and $(1-q_0)\frac{1}{J}$ to the corresponding J controls to form case-control weighted estimates of $E^*_0(Y_1) = P^*_0(Y_1=1)$ and $E^*_0(Y_0) = P^*_0(Y_0=1)$. The risk difference, relative risk, and odds ratio can then be simply calculated from these estimates. For example, the relative risk $E^*_0(Y_1)/E^*_0(Y_0)$ is estimated by:

$$\hat{}_{RR} \;=\; \frac{\frac{1}{n}\sum_{i=1}^{n} q_0 \hat{Q}^*_{1,q_0}(1, W_{1i}) + (1-q_0)\frac{1}{J}\sum_j \hat{Q}^*_{1,q_0}(1, W^j_{2i})}{\frac{1}{n}\sum_{i=1}^{n} q_0 \hat{Q}^*_{1,q_0}(0, W_{1i}) + (1-q_0)\frac{1}{J}\sum_j \hat{Q}^*_{1,q_0}(0, W^j_{2i})}.$$

3.2.6 Calculating Standard Errors

The calculation of standard errors for case-control weighted targeted maximum likelihood involves the use of case-control weighted influence curves for the risk difference, relative risk, and odds ratio. This methodology is discussed in detail in van der Laan (2008), and a complete technical understanding of infuence curve derivation is not necessary to implement the case-control targeted maximum likelihood estimation procedure. We also refer to van der

Laan and Robins (2002) for careful discussions of gradients and influence curve theory.

For example, the unweighted influence curve for the risk difference of a prospective study $\psi^*_{0,RD} = P^*_0(Y_1 = 1) - P^*_0(Y_0 = 1)$ is estimated by:

$$\begin{aligned}\hat{D}_{RD}(\psi^*, g^*, Q^*)(O) &= \frac{I(A=1)}{\hat{g}^*(1 \mid W)}(Y - \hat{Q}^*(1,W)) - \frac{I(A=0)}{\hat{g}^*(0 \mid W)}(Y - \hat{Q}^*(0,W)) \\ &\quad + \hat{Q}^*(1,W) - \hat{Q}^*(0,W) - \hat{\psi}.\end{aligned}$$

The case-control weighted influence curve for the risk difference $\psi^*_{0,RD} = P^*_0(Y_1 = 1) - P^*_0(Y_0 = 1)$ is then estimated by:

$$\begin{aligned}\hat{D}_{RD,q_0}(\psi^*, g^*, Q^*)(O) &= q_0 \hat{D}^*(g^*, Q^*)(A_1, W_1, 1) \\ &\quad + (1 - q_0)\frac{1}{J}\sum_{j=1}^{J} \hat{D}^*(g^*, Q^*)(A^j_2, W^j_2, 0) - \hat{\psi}.\end{aligned}$$

Note that the case-control weighted influence curve is merely the influence curve for prospective targeted maximum likelihood with case-control weighting. See van der Laan and Rubin (2006) and Moore and van der Laan (2007) for prospective sampling targeted maximum likelihood methodology.

An estimate of the asymptotic variance of $\sqrt{n}(\hat{\psi} - \psi^*_0)$ using the estimate of the efficient influence curve $D_{q_0}(\psi^*, g^*, Q^*)(O)$ is given by:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} D^2_{q_0}(\psi^*, g^*, Q^*)(O).$$

Given the influence curve for the causal parameter estimate $\hat{\psi}$, a 95% Wald-type confidence interval can be constructed as: $\hat{\psi} \pm z_{0.975}\frac{\hat{\sigma}}{\sqrt{n}}$. Likewise, the p-value of $\hat{\psi}$ can be calculated as $2[1 - \Phi(|\frac{\hat{}}{\hat{\sigma}/\sqrt{n}}|)]$.

4 Intercept Adjusted MLE and Case-Control Weighted MLE

Intercept adjusted maximum likelihood estimation and case-control weighted maximum likelihood estimation were previously discussed as options for the initial fit $\hat{Q}^*(A, W)$. Several issues became apparent when using intercept adjusted maximum likelihood estimation for $\hat{Q}^*(A, W)$ in our case-control weighted targeted maximum likelihood framework. In multiple simulation settings we found that when $\hat{Q}^*(A, W)$ was misspecified using an intercept adjusted fit, the predicted probabilities were substantially biased compared to the misspecified case-control weighted maximum likelihood probabilties. This additional bias can be understood intuitively since the update to the logistic regression $\log c_0$ is static regardless of the model used, and the parameters of the model (excluding the intercept) are not adjusted by this update. For correctly specified $\hat{Q}^*(A,W)$ this is not an issue, but when $\hat{Q}^*(A,W)$ is misspecified, it leads to substantial bias. Conversely, the case-control weighted logistic regression estimate incorporates the case-control weights each time it fits an estimate. Thus, for misspecified $\hat{Q}^*(A,W)$, case-control weighted predicted probabilities will likely be closer to the truth than intercept adjusted estimates. See Figure 1 for an illustration.

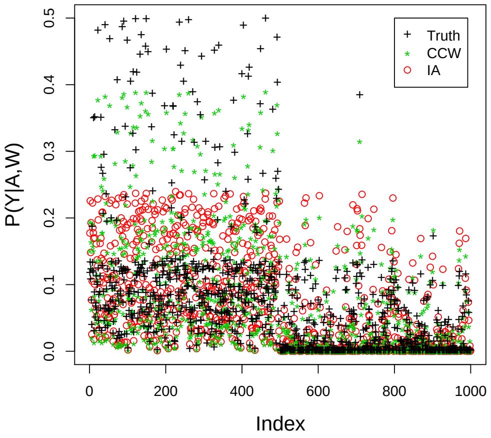


Figure 1: **Predicted Probabilities for Misspecified $\hat{Q}^*(A,W)$.**

The case-control targeted intercept adjusted maximum likelihood improved, with regard to bias, on its non-targeted counterpart for misspecified $\hat{Q}^*(A,W)$. However, the additional bias for misspecified $\hat{Q}^*(A,W)$ and intercept adjusted logistic regression led to much slower convergence to the true values of the risk difference, relative risk, and odds ratio within the case-control targeted maximum likelihood framework. Case-control weighted targeted maximum likelihood with misspecified $\hat{Q}^*(A,W)$ fit with case-control weighted logistic regression became consistent for reasonable sample sizes. Coverage probabilities for case-control weighted targeted intercept adjusted maximum likelihood estimation for misspecified $\hat{Q}^*(A,W)$ also diverged substantially from 95% (as low as 65%) for reasonable sample sizes due to the bias of the estimators. We should note that when $\hat{Q}^*(A,W)$ is correctly specified, the intercept adjusted methods performed as well as the case-control weighted methods. However, the correct specification of $\hat{Q}^*(A, W)$ is unlikely in practice. Given these findings, we present in our simulations the use of case-control weighted targeted maximum likelihood estimation using case-control weighted logistic regression for the initial fit.

5 Simulation Studies

5.1 Simulation 1

Our first simulation study was designed to illustrate the advantages of the case-control weighting scheme for targeted maximum likelihood estimation in case-control designs. It was based on a population of $N = 120,000$ individuals, where we simulated a 1-dimensional covariate W, a binary exposure A, and indicator Y, which was 1 for cases and 0 for controls. These variables were generated according to the following rules:

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

$$g_0^*(A \mid W) = \frac{1}{1+\exp(-(W^2-4W+1))}$$

$$Q_0^*(A,W) = \frac{1}{1+\exp(-(1.2A-\sin(W^2)+A\sin(W^2)+5A\log(W)+5\log(W)-1))}.$$

The resulting population had a prevalence probability $q_0 = 0.035$, and exactly $4,165$ cases. We sampled the population using a varying number of cases and controls, and for each sample size we ran 1000 simulations. The true values of the risk difference, relative risk, and odds ratio were given by $RD = 0.043$, $RR = 2.483$, and $OR = 2.598$, with $P(Y_1 = 1) = 0.072$ and $P(Y_0 = 1) = 0.029$. These causal effect parameters were estimated using methods discussed in this paper:

1. **IPTW**: IPTW method for marginal structural models (Robins, 1999; Mansson et al., 2007) that uses the estimated exposure mechanism among the controls to update a logistic regression of Y on A discussed in Section 2.2.

2. **Case-Control Weighted MLE (CCW-MLE)**: Case-control weighted logistic regression, discussed in Section 3.2.1, mapped to causal effect estimators by averaging over the case-control weighted distribution of W.

3. **Case-Control Weighted Targeted MLE (CCW-TMLE)**: Case-control weighted targeted maximum likelihood procedure for case-control designs with case-control weighted $\hat{Q}^*(A,W)$ discussed in Section 3.

The initial fit for each method requiring an estimate of $Q_0^*(A,W)$ was defined by:

$$\hat{Q}^*(A,W) = \frac{1}{1+\exp(-(\hat{\alpha_0}+\hat{\alpha_1}A+\hat{\alpha_2}\log(W)+\hat{\alpha_3}\sin(W^2)+\hat{\alpha_4}A\log(W)+\hat{\alpha_5}A\sin(W^2)))},$$

which was the correctly specified fit. $Q_0^*(A,W)$ was also estimated in a second simulation with:

$$\hat{Q}^*(A,W) = \frac{1}{1+\exp(-(\hat{\alpha_0}+\hat{\alpha_1}A+\hat{\alpha_2}W))},$$

a misspecified fit. For methods requiring a fit for exposure mechanism, the correct fit was defined by:

$$\hat{g}^*(A \mid W) = \frac{1}{1+\exp(\hat{\eta_0}+\hat{\eta_1}W^2+\hat{\eta_2}W)}.$$

The misspecified version of the exposure mechanism was given by:

$$\hat{g}^*(A \mid W) = \frac{1}{1+\exp(\hat{\eta_0}+\hat{\eta_1}W)}.$$

In our simulation study, we realistically generated A dependent on W. This led to some substantial increases in efficiency in the targeted estimator when $\hat{Q}^*(A,W)$ was misspecified and sample size was larger, as they also adjust for $\hat{g}^*(A \mid W)$. This emphasizes the double robustness of the targeted estimators, and suggests that one should always adjust for $\hat{g}^*(A \mid W)$ in practice. When $\hat{Q}^*(A,W)$ was correctly specified, the relative efficiency of the targeted estimator (CCW-TMLE) was similar to its non-targeted counterpart (CCW-MLE), demonstrating that the use of q_0 and $\hat{Q}^*(A,W)$ alone can produce efficient estimators. This was further highlighted in the results for the odds ratio and the IPTW estimators, which do not utilize q_0, as they had the poorest overall efficiency. Mean squared errors and relative efficiencies for the causal odds ratio are provided in Table 1. The results for the relative risk and risk difference are combined in Table 2. The least efficient estimator as sample size increased for these parameters was the case-control weighted logistic regression when $Q_0^*(A,W)$ was realistically misspecified.

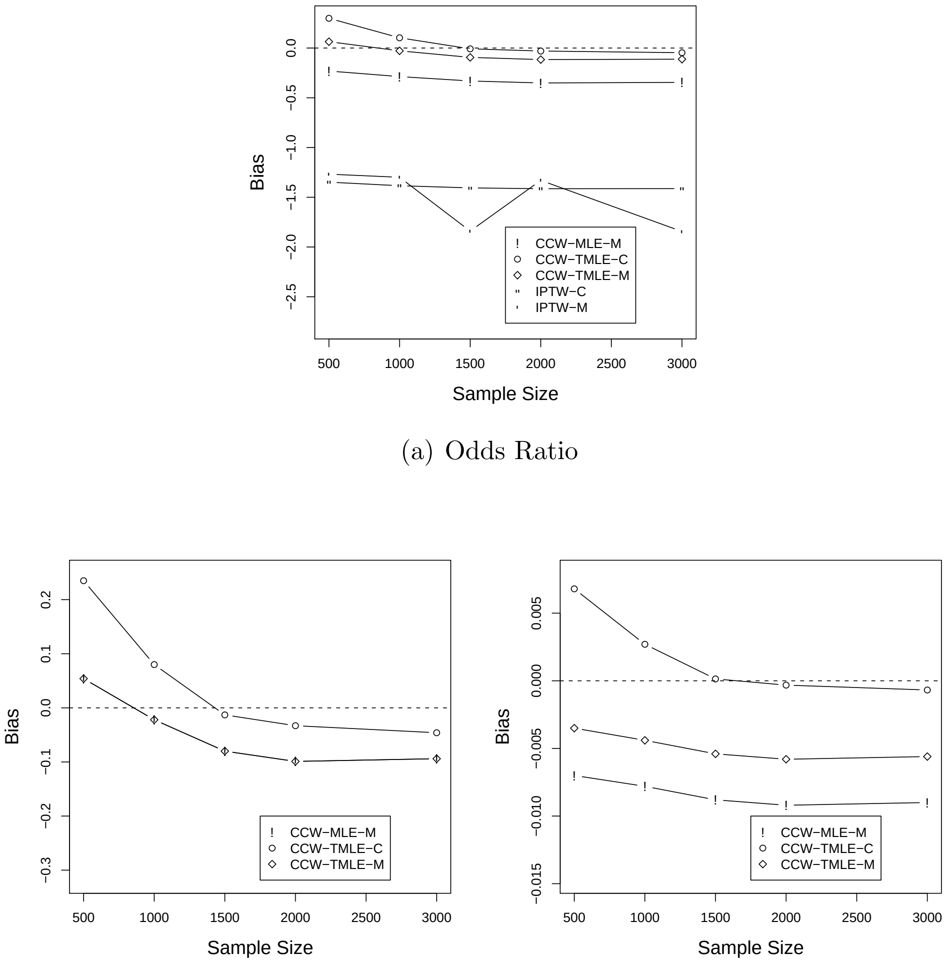
When examining the bias of the estimators for the odds ratio, it is clear that the IPTW estimators had the highest level of bias across all sample sizes,

Table 1: **Simulation 1 – Odds Ratio** – MSE is Mean Squared Error for IPTW Misspecified Estimate, RE is Relative Efficiency of Other Estimators Compared to IPTW Misspecified Estimate MSE, nC is Number of Cases, nCo is Number of Controls, n is Number of Total Observations, M is for Misspecified $\hat{Q}^*(A,W)$ or $\hat{g}^*(A \mid W)$ Fit, C is for Correctly Specified $\hat{Q}^*(A,W)$ or $\hat{g}^*(A \mid W)$. (When two letters are noted in the "Fit" column, the first letter refers to $\hat{Q}^*(A,W)$ and the second to $\hat{g}^*(A \mid W)$.)

		n=500	n=1000	n=1500	n=2000	n=3000
		nC=250	nC=500	nC=500	nC=1000	nC=1000
Odds Ratio	Fit	nCo=250	nCo=500	nCo=1000	nCo=1000	nCo=2000
IPTW MSE	M	1.76	1.75	3.39	1.80	3.40
IPTW RE	C	0.91	0.89	1.69	0.89	1.69
	C/C	1.27	3.62	14.58	8.40	32.03
CCW-TMLE RE	C/M	1.26	3.62	14.57	8.40	31.97
	M/C	1.96	4.63	16.68	9.52	31.91
CCW-MLE RE	C	1.27	3.65	14.64	8.44	32.12
	M	3.07	5.72	14.54	7.83	18.93

Table 2: **Simulation 1 – Relative Risk and Risk Difference** – MSE is Mean Squared Error for CCW-MLE Misspecified Estimate, RE is Relative Efficiency of Other Estimators Compared to CCW-MLE Misspecified Estimate MSE, nC is Number of Cases, nCo is Number of Controls, n is Number of Total Observations, M is for Misspecified $\hat{Q}^*(A,W)$ or $\hat{g}^*(A \mid W)$ Fit, C is for Correctly Specified $\hat{Q}^*(A,W)$ or $\hat{g}^*(A \mid W)$. (When two letters are noted in the "Fit" column, the first letter refers to $\hat{Q}^*(A,W)$ and the second to $\hat{g}^*(A \mid W)$.)

		n=500	n=1000	n=1500	n=2000	n=3000
		nC=250	nC=500	nC=500	nC=1000	nC=1000
Relative Risk	Fit	nCo=250	nCo=500	nCo=1000	nCo=1000	nCo=2000
CCW-MLE MSE	M	0.46	0.25	0.19	0.19	0.15
CCW-MLE RE	C	0.48	0.69	1.06	1.12	1.73
	C/C	0.47	0.68	1.05	1.12	1.73
CCW-TMLE RE	C/M	0.47	0.68	1.05	1.12	1.73
	M/C	0.65	0.82	1.15	1.22	1.69
Risk Difference						
CCW-MLE MSE	M	3.2E-04	1.8E-04	1.4E-04	1.4E-04	1.1E-04
CCW-MLE RE	C	0.45	0.67	1.10	1.15	1.89
	C/C	0.45	0.67	1.10	1.15	1.89
CCW-TMLE RE	C/M	0.45	0.67	1.10	1.15	1.89
	M/C	0.98	1.12	1.34	1.36	1.63



(a) Odds Ratio

(b) Relative Risk

(c) Risk Difference

Figure 2: **Simulation 1 – Bias Results.** (Bias results for the case-control weighted targeted maximum likelihood with misspecified $\hat{g}^*(A \mid W)$ and the correctly specified case-control weighted targeted maximum likelihood were excluded since those values were the same as those for the targeted maximum likelihood with correctly specified $\hat{Q}^*(A, W)$ and $\hat{g}^*(A \mid W)$.)

as observed in the bias plot displayed in Figure 2(a). The case-control weighted logistic regression and case-control weighted targeted maximum likelihood with misspecified $\hat{Q}^*(A, W)$ had more bias than their correctly specified counterparts. It may be possible to avoid some of the additional bias caused by the misspecification of $\hat{Q}^*(A, W)$ in practice by fitting $\hat{Q}^*(A, W)$ with data-adaptive model-selection, such as the Deletion/Substitution/Addition (DSA) algorithm or other readily available machine learning algorithms. For more details about this procedure we refer to Sinisi and van der Laan (2004). The bias results for the relative risk and risk difference followed similar trends, as can be seen in Figure 2(b) and 2(c). While the case-control weighted logistic regression has low variance when misspecified, it may be more biased than its targeted counterpart. These results bolster our theoretical arguments that gains in efficiency and reduction in bias can be obtained by having a known prevalence probability and using a targeted estimator. Additionally, under typical circumstances experienced in an experimental setting, the case-control weighted targeted maximum likelihood may perform the best with regard to bias and variability.

5.2 Simulation 2

Our second set of simulations was based on a population of $N = 80,000$ individuals, and was designed to illustrate, in another setting, the advantages of incorporating known prevalence probability into case-control design methodology. The population was generated with binary exposure A and disease status Y and a 1-dimensional covariate W. These variables were generated according to the following rules:

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

$$g_0^*(A \mid W) = P_0^*(A = 1|W) = \tfrac{1}{1+\exp(-5\sin(W)))}$$

$$Q_0^*(A, W) = P_0^*(Y = 1|A, W) = \tfrac{1}{1+\exp(-(2A-25W+AW))}.$$

The resulting population had a prevalence probability $q_0 = 0.053$, exactly $4,206$ cases, and also followed an independent case-control sampling design. The true values of the risk difference, relative risk, and odds ratio were given by $RD = 0.061$, $RR = 3.21$, and $OR = 3.42$, with $P(Y_1 = 1) = 0.089$ and $P(Y_0 = 1) = 0.028$. These parameters were estimated using the same general methods as in the previous section, albeit with different fits for $\hat{Q}^*(A, W)$ and $\hat{g}^*(A \mid W)$. The initial fit for each method requiring a fit for $\hat{Q}^*(A, W)$ was

Table 3: **Simulation 2 – Odds Ratio** – MSE is Mean Squared Error for IPTW misspecified Estimate, RE is Relative Efficiency of Other Estimators Compared to IPTW misspecified Estimate MSE, nC is Number of Cases, nCo is Number of Controls, n is Number of Total Observations, M is for Misspecified $\hat{Q}^*(A, W)$ or $\hat{g}^*(A \mid W)$ Fit, C is for Correctly Specified $\hat{Q}^*(A, W)$ or $\hat{g}^*(A \mid W)$. (When two letters are noted in the "Fit" column, the first letter refers to $\hat{Q}^*(A, W)$ and the second to $\hat{g}^*(A \mid W)$.)

		n=350	n=500	n=750	n=1000
		nC=100	nC=250	nC=250	nC=500
Odds Ratio	Fit	nCo=250	nCo=250	nCo=500	nCo=500
IPTW MSE	M	404.40	3667.56	306.42	2433.62
IPTW RE	C	1.0E+00	1.2E+00	1.0E+00	1.2E+00
CCW-TMLE RE	C/C	2.8E+02	4.1E+03	5.7E+02	5.7E+03
	C/M	2.9E+02	4.1E+03	5.7E+02	5.7E+03
CCW-MLE RE	C	2.9E+02	4.2E+03	5.7E+02	5.8E+03

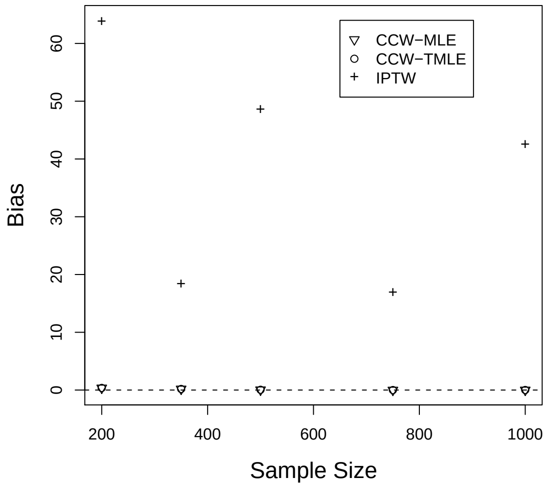


Figure 3: **Simulation 2 – Bias Results for the Odds Ratio.** (Bias results for the case-control weighted targeted maximum likelihood with misspecified $\hat{g}^*(A \mid W)$ were excluded since those values were the same as those for the targeted maximum likelihood with correctly specified $\hat{Q}^*(A, W)$ and $\hat{g}^*(A \mid W)$.)

defined by:

$$\hat{Q}^*(A,W) = \frac{1}{1+\exp(-(\hat{\alpha_0}+\hat{\alpha_1}A+\hat{\alpha_2}W+\hat{\alpha_3}AW))},$$

which was the correctly specified fit. For methods requiring a fit for exposure mechanism, the correct fit was defined by:

$$\hat{g}^*(A \mid W) = \frac{1}{1+\exp(-(\hat{\eta_0}+\hat{\eta_1}\sin(W)))}.$$

The misspecified version of the exposure mechanism was given by:

$$\hat{g}^*(A \mid W) = \frac{1}{1+\exp(-(\hat{\eta_0}+\hat{\eta_1}W))}.$$

Results across the two case-control weighted methods for the risk difference, relative risk, and odds ratio were nearly identical, indicating in this example that when $\hat{Q}^*(A,W)$ is correct and q_0 is known, one may be well served by either of these methods. However, the IPTW method for odds ratio estimation was quite inefficient in comparison. We theorized in van der Laan (2008), and Mansson et al. (2007) demonstrated, that the IPTW procedure has a strong sensitivity towards model misspecification. This result was seen in Simulation 1, although the results in Simulation 2 are more extreme. Results for the odds ratio estimation can be seen in Table 3 and Figure 3. Again we see that gains in efficiency and reduction in bias can be obtained by simply having known q_0.

5.3 Standard Errors, Confidence Intervals, and P-Values

Continuing with the simulated population from Simulation 2, we provide an example of the use of influence curves in the estimation of standard errors for case-control weighted targeted maximum likelihood estimation. We sampled one data set of $n = 1000$ from the population, with equal numbers of cases and controls, and estimated the odds ratio. Recall that the true value for the odds ratio was given by $OR = 3.42$. The case-control weighted targeted maximum likelihood estimator uses the influence curve to estimate standard errors, as discussed in Section 3.2.6, with estimated variance given by $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} D_{q_0}^2(\psi^*, g^*, Q^*)(O)$. Standard error estimates for the IPTW estimator were calculated by bootstrapping the case and control samples 1000 times. The results for this single sampling of the simulated population can be seen in Table 4, including odds ratio estimates, standard errors, confidence intervals, and p-values. It compares only the case-control weighted targeted

Table 4: **Standard Error Illustration** – OR is Odds Ratio Estimate, SE is Standard Error, CI is Confidence Interval, P is P-value, C is for Correctly Specified $\hat{Q}^*(A, W)$ or $\hat{g}^*(A \mid W)$, M is for Misspecified $\hat{g}^*(A \mid W)$. (When two letters are noted in the "Fit" column, the first letter refers to $\hat{Q}^*(A, W)$ and the second to $\hat{g}^*(A \mid W)$.) The results are for one data set of 1000 individuals with 500 cases and 500 controls randomly sampled from the population in Simulation 2. True $OR = 3.42$.

Odds Ratio	Fit	OR	SE	CI	P
IPTW	C	64.98	22.44	[21.00, 108.96]	0.004
	M	64.64	4.66	[55.50, 73.77]	< 0.001
CCW-TMLE RE	C/C	3.39	0.24	[2.93, 3.85]	< 0.001
	C/M	3.39	0.24	[2.92, 3.86]	< 0.001

maximum likelihood estimator and the IPTW estimator. (The non-targeted maximum likelihood method was excluded as we wish to draw attention to the use of the influence curve for standard error estimation. Standard errors for the non-targeted maximum likelihood method can also be calculated using bootstrapping.)

5.4 Simulation 3

Our third simulation study was designed to illustrate the performance of the case-control weighting scheme for targeted maximum likelihood estimation in case-control designs when q_0 is estimated. We also examine coverage probabilities and percentage of rejected tests for case-control weighted targeted maximum likelihood estimation. The simulation was based on a population of $N = 120,000$ individuals, and we simulated a 1-dimensional covariate W, binary exposure A, and indicator Y. The variables were generated according to the following rules:

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

$$g_0^*(A \mid W) = P_0^*(A = 1|W) = \frac{1}{1+\exp(-(W^2-4W+1))}$$

$$Q_0^*(A, W) = P_0^*(Y = 1|A, W) = \frac{1}{1+\exp(-(A-\sin(W^2)+A\sin(W^2)+7A\log(W)+5\log(W)-1))}.$$

The resulting population had a prevalence probability $q_0 = 0.032$, and exactly $3,834$ cases. We ran 1000 simulations and sampled 500 cases and 500 controls for varying levels of the prevalence probability $q_0 = (0.02, 0.03, 0.04)$. The true value for the odds ratio was given by $OR = 1.851$, with $P(Y_1 = 1) = 0.052$ and $P(Y_0 = 1) = 0.029$. The causal odds ratio was estimated using case-control weighted targeted maximum likelihood estimation. The correctly specified initial fit for $Q_0^*(A, W)$ was estimated by:

$$\hat{Q}^*(A, W) = \frac{1}{1+\exp(-(\hat{\alpha_0}+\hat{\alpha_1}A+\hat{\alpha_2}\log(W)+\hat{\alpha_3}\sin(W^2)+\hat{\alpha_4}A\log(W)+\hat{\alpha_5}A\sin(W^2)))}.$$

The misspecified initial fit was estimated with:

$$\hat{Q}^*(A, W) = \frac{1}{1+\exp(-(\hat{\alpha_0}+\hat{\alpha_1}A+\hat{\alpha_2}W))}.$$

For exposure mechanism, the correct fit was defined by:

$$\hat{g}^*(A \mid W) = \frac{1}{1+\exp(\hat{\eta_0}+\hat{\eta_1}W^2+\hat{\eta_2}W)}.$$

The misspecified version of the exposure mechanism was given by:

$$\hat{g}^*(A \mid W) = \frac{1}{1+\exp(\hat{\eta_0}+\hat{\eta_1}W)}.$$

When examining the mean squared error results of the odds ratio across the range of values for q_0, one can see deviations away from the values obtained for the true q_0. However, it is important to note that the coverage probabilities (the percentage of simulations where the estimated confidence interval contained the true odds ratio) were not highly variant and remain near 95%. This provides evidence that the case-control weighted targeted maximum likelihood procedure performs well with estimated values of q_0. The percentage of rejected tests ($\alpha = 0.05$) across the range of q_0 was also relatively stable. The results for the mean squared errors, coverage probabilities, and percent rejected tests for the odds ratio can be seen in Table 5. Simulations that resample q_0 from its sampling distribution could also be used to get an estimate of the total uncertainty surrounding the parameter of interest, but they are not explored here. An analytic equivalent to this resampling can be found in the appendix to van der Laan (2008). This theorem demonstrates that one can incorporate the standard error of the estimate $\hat{q}_0$ into the confidence interval for the parameter of interest.

Table 5: **Simulation 3 – Odds Ratio** – MSE is Mean Squared Error, CP is Coverage Probability (percentage of simulations where estimated confidence interval contained the true odds ratio), Rej is for Percent Rejected Tests ($\alpha = 0.05$), C is for Correctly Specified $\hat{Q}^*(A,W)$ or $\hat{g}^*(A \mid W)$, M is for Misspecified $\hat{Q}^*(A,W)$ or $\hat{g}^*(A \mid W)$. (When two letters are noted in the "Fit" column, the first letter refers to $\hat{Q}^*(A,W)$ and the second to $\hat{g}^*(A \mid W)$.) The results are for 1000 simulations of 1000 individuals with 500 cases and 500 controls randomly sampled from the population in Simulation 3. True $OR = 1.851$. True $q_0 = 0.032$.

		True q_0		q_0	
	Fit	0.032	0.020	0.030	0.040
	C/C	0.35	0.74	0.39	0.24
CCW-TMLE MSE	C/M	0.35	0.74	0.39	0.24
	M/C	0.19	0.28	0.20	0.16
	C/C	0.94	0.95	0.94	0.92
CCW-TMLE CP	C/M	0.97	0.97	0.97	0.95
	M/C	0.92	0.94	0.93	0.91
	C/C	0.33	0.32	0.33	0.34
CCW-TMLE Rej	C/M	0.21	0.23	0.22	0.20
	M/C	0.02	0.01	0.02	0.03

6 Discussion

Case-control weighted targeted maximum likelihood estimation provides a framework for the analysis of case-control study designs. We observed that the IPTW method for causal parameter estimation was outperformed in conditions similar to a practical setting by the new case-control weighted targeted maximum likelihood estimation methodology. The case-control weigted targeted maximum likelihood estimation procedure yields a fully robust and locally efficient estimator of several marginal causal parameters of interest. Model misspecification within this framework, with known exposure mechanism, still results in efficient estimatiors. Additionally, the case-control weighted logistic regression mapped to causal parameters had high efficiency and reduced bias in comparison to the IPTW estimator. This is an important result for those applied researchers who may not feel comfortable implementing the case-control weighted targeted maximum likelihood procedure. Thus, we showed striking improvements in efficiency and bias in all methods incorporating knowledge of the prevalence probability over the IPTW estimator which does not use this information. Knowledge of the prevalence probability may be realistic in many settings. Where possible, researchers might consider prioritizing accurately defining their population of interest, which will streamline obtaining or estimating the prevalence probability. We also demonstrated that a range of values for q_0 can be used with case-control weighted targeted maximum likelihood estimation to obtain efficient causal parameters of interest. As addressed earlier, we discussed case-control weighted targeted maximum likelihood estimation for cumulative study designs with the prevalence probability. Future areas of work include adapting our methods for density sampling, where controls are drawn from the population at risk at the time a case develops disease. For example, using case-control weights that depend on the time points the cases and controls were sampled, as discussed in an appendix in van der Laan (2008). Here, the use of incidence probabilities would be more appropriate.

References

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

O. Bembom, M.L. Peterson, S-Y Rhee, W.J. Fessel, S.E. Sinisi, R.W. Shafer, and M.J. van der Laan. Biomarker discovery using targeted maximum likelihood estimation: Application to the treatment of antiretroviral resistant hiv infection. *Technical Report 221, Division of Biostatistics, University of California, Berkeley*, 2007.

J. Benichou and S. Wacholder. A comparison of three approaches to estimate exposure-specific incidence rates from population-based case-control data. *Statistics in Medicine*, 13:651–661, 1994.

N.E. Breslow. Statistics in epidemiology: the case-control study. *J Am Stat Soc*, 91:14–28, 1996.

N.E. Breslow and N.E. Day. *Statistical Methods in Cancer Research: Volume 1 – The analysis of case-control studies.* International Agency for Research on Cancer, Lyon, 1980.

J. Cornfield. A method of estimating comparative rates from clinical data. applications to cancer of the lung, breast, and cervix. *J Nat Cancer Inst*, 11:1269–1275, 1951.

J. Cornfield. A statistical problem arising from retrospective studies. In J. Neyman, editor, *Proceedings of the Third Berkeley Symposium, Volume IV*, pages 133–148. University of California Press, 1956.

S. Greenland. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *Am J Epidemiol*, 160(4):301–305, 2004.

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

P.W. Holland and D.B. Rubin. Causal inference in retrospective studies. In D.B. Rubin, editor, *Matched Sampling for Causal Effects*. Cambridge University Press, Cambridge, MA, 1988.

R. Mansson, M.M. Joffe, W Sun, and S. Hennessy. On the estimation and use of propensity scores in case-control and case-cohort studies. *Am J Epidemiol*, 166(3):332–339, 2007.

K.L. Moore and M.J. van der Laan. Covariate adjustment in randomized trials with binary outcomes. *Technical Report 215, Division of Biostatistics, University of California, Berkeley*, 2007.

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.

S. Newman. Causal analysis of case-control data. *Epid Persp Innov*, 3:2, 2006. URL `http://www.epi-perspectives.com/content/3/1/2`.

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.

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

S. Rose and M.J. van der Laan. Why match? investigating matched case-control study designs with causal effect estimation. *Technical Report 240, Division of Biostatistics, University of California, Berkeley*, 2008.

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

S. Sinisi and M.J. van der Laan. Deletion/substitution/addition algorithm in loss function based estimation. *Journal of Statistical Methods in Molecular Biology*, 3(1):Article 18, 2004.

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 incidence 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 causailty.* Springer, New York, 2002.

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

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