Chapter 13
Independent Case-Control Studies

Sherri Rose, Mark J. van der Laan

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, as they are able to sample known cases of disease vs. following a large number of subjects and waiting for disease onset in a relatively small number of individuals.

Case-control sampling is a biased design. Bias occurs due to the disproportionate number of cases in the sample vs. the population.

Researchers commonly employ the use of logistic regression in a parametric statistical model, ignoring the biased design, and estimate the conditional odds ratio of having disease given the exposure of interest A and measured covariates W.

Our proposed case-control-weighted TMLE for case-control studies relies on knowledge of the true prevalence probability, or a reasonable estimate of this probability, to eliminate the bias of the case-control sampling design. We use the prevalence probability in case-control weights, and our case-control weighting scheme successfully maps the TMLE for a random sample into a method for case-control sampling. The case-control-weighted TMLE (CCW-TMLE) is an efficient estimator for the case-control sample when the TMLE for the random sample is efficient. In addition, the CCW-TMLE inherits the robustness properties of the TMLE for the random sample.

13.1 Data, Model, and Target Parameter

Let us define a simple example with $X = (W, A, Y) \sim P_{X,0}$ as the full-data experimental unit and corresponding distribution $P_{X,0}$ of interest, which consists of baseline covariates W, exposure variable A, and a binary outcome Y that defines case or

M.J. van der Laan and S. Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer Series in Statistics, DOI 10.1007/978-1-4419-9782-1_13,
© Springer Science+Business Media, LLC 2011

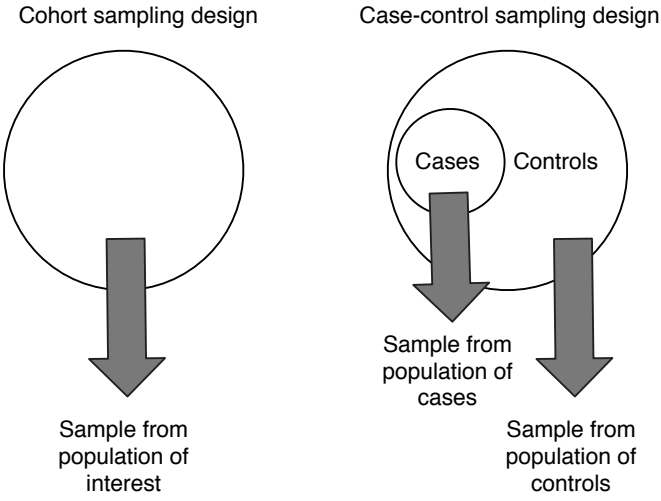

Fig. 13.1 Case-control sampling design

control status. In previous chapters, our target parameter of interest was the causal risk difference, which we now denote

$$\begin{aligned}\psi^F_{RD,0} = \Psi^F(P_{X,0}) &= E_{X,0}[E_{X,0}(Y \mid A = 1, W) - E_{X,0}(Y \mid A = 0, W)] \\ &= E_{X,0}(Y_1) - E_{X,0}(Y_0) \\ &= P_{X,0}(Y_1 = 1) - P_{X,0}(Y_0 = 1)\end{aligned}$$

for binary A, binary Y, and counterfactual outcomes Y_0 and Y_1, where F indicates "full data." Other common parameters of interest include the causal relative risk and the causal odds ratio, given by

$$\psi^F_{RR,0} = \frac{P_{X,0}(Y_1 = 1)}{P_{X,0}(Y_0 = 1)}$$

and

$$\psi^F_{OR,0} = \frac{P_{X,0}(Y_1 = 1)P_{X,0}(Y_0 = 0)}{P_{X,0}(Y_1 = 0)P_{X,0}(Y_0 = 1)}.$$

We describe the case-control design as first sampling (W_1, A_1) from the conditional distribution of (W, A), given $Y = 1$ for a case. One then samples J controls (W_0^j, A_0^j) from (W, A), given $Y = 0, j = 1, \ldots, J$. The observed data structure in independent case-control sampling is then defined by

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

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

where the cluster containing one case and J controls is considered the experimental unit. Therefore, a case-control data set consists of n independent and identically distributed observations $O_1, \ldots, O_n$ with sampling distribution P_0 as described above. The statistical model $\mathcal{M}^F$, where the prevalence probability $P_{X,0}(Y = 1) \equiv q_0$ may or may not be known, implies a statistical model for the distribution of O consisting of (W_1, A_1) and controls (W_2^j, A_2^j), $j = 1, \ldots, J$.

This coupling formulation is useful when proving theoretical results for the case-control weighting methodology (van der Laan 2008a), and those results 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. We also stress that this formulation does not describe *individually matched* case-control sampling, which we describe in Chap. 14.

13.2 Prevalence Probability

The population under study should be clearly defined. As such, the prevalence probability q_0 is then truly basic information about a population of interest. The use of the prevalence probability to eliminate the bias of a case-control sampling design as an update to a logistic regression intercept in a parametric statistical model was first discussed in Anderson (1972). This update enforces the intercept to be equal to $\log(q_0/(1 - q_0))$.

13.3 CCW-TMLE

In this section we build on the readers familiarity with the TMLE as described in detail in Chaps. 4 and 5. We discuss a CCW-TMLE for the causal risk difference with $X = (W, A, Y) \sim P_{X,0}$ and $O = \left((W_1, A_1), (W_0^j, A_0^j : j = 1, \ldots, J)\right) \sim P_0$. The full-data efficient influence curve $D^F(Q_0, g_0)$ at $P_{X,0}$ is given by

$$D^F(Q_0, g_0) = \left(\frac{I(A = 1)}{g_0(1 \mid W)} - \frac{I(A = 0)}{g_0(0 \mid W)}\right)(Y - \bar{Q}_0(A, W)) + \bar{Q}_0(1, W) - \bar{Q}_0(0, W) - \Psi^F(Q_0), \tag{13.1}$$

where $Q_0 = (\bar{Q}_0, Q_{W,0})$, $Q_{W,0}$ is the true full-data marginal distribution of W, $\bar{Q}_0(A, W) = E_{X,0}(Y \mid A, W)$, and $g_0(a \mid W) = P_{X,0}(A = a \mid W)$. The first term will be denoted by D_Y^F and the second term by D_W^F, since these two terms represent

components of the full-data efficient influence curve that are elements of the tangent space of the conditional distribution of Y, given (A, W), and the marginal distribution of W, respectively. That is, D_Y^F is the component of the efficient influence curve that equals a score of a parametric fluctuation model of a conditional distribution of Y, given (A, W), and D_W^F is a score of a parametric fluctuation model of the marginal distribution of W. Note that $D_Y^F(Q, g)$ equals a function $H^*(A, W)$ times the residual $(Y - \bar{Q}(A, W))$, where

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

13.3.1 Case-Control-Weighted Estimators for Q_0 and g_0

We can estimate the marginal distribution of $Q_{W,0}$ with case-control-weighted maximum likelihood estimation:

$$Q_{W,n}^0 = \arg\min_{Q_W} \sum_{i=1}^{n} \left(q_0 L^F(Q_W)(W_{1,i}) + \frac{1 - q_0}{J} \sum_{j=1}^{J} L^F(Q_W)(W_{2,i}^j) \right),$$

where $L^F(Q_W) = -\log Q_W$ is the log-likelihood loss function for the marginal distribution of W. If we maximize over all distributions, this results in a case-control-weighted empirical distribution that puts mass q_0/n on the cases and $(1 - q_0)/(nJ)$ on the controls in the sample.

Suppose that based on a sample of n i.i.d. observations X_i we would have estimated $\bar{Q}_0$ with loss-based learning using the log-likelihood loss function $L^F(\bar{Q})(X) = -\log \bar{Q}(A, W)^Y (1 - \bar{Q}(A, W))^{1-Y}$. Given the actual observed data we can estimate $\bar{Q}_0$ with super learning and the case-control weights for observations $i = 1, \ldots, n$, which corresponds with the same super learner but now based on the case-control-weighted loss function:

$$L(\bar{Q})(O) \equiv q_0 L^F(\bar{Q})(W_1, A_1, 1) + \frac{1 - q_0}{J} \sum_{j=1}^{J} L^F(\bar{Q})(W_2^j, A_2^j, 0).$$

Let $L^F(Q) = L^F(Q_W) + L^F(\bar{Q})$ be the full-data loss function for $Q = (\bar{Q}, Q_W)$, and let $L(Q, q_0) = q_0 L^F(Q)(W_1, A_1, 1) + ((1 - q_0)/J) \sum_{j=1}^{J} L^F(Q)(W_2^j, A_2^j, 0)$ be the corresponding case-control-weighted loss function. We have $Q_0 = \arg\min_Q E_{P_0} L(Q, q_0)(O)$, so that indeed the case-control-weighted loss function for Q_0 is a valid loss function. Similarly, we can estimate g_0 with loss-based super learning based on the case-control-weighted log-likelihood loss function:

$$L(g)(O) \equiv -q_0 \log g(A_1 \mid W_1) - \frac{1 - q_0}{J} \sum_{j=1}^{J} \log g(A_2^j \mid W_2^j).$$

We now have an initial estimator $Q_n^0 = (Q_{W,n}^0, \bar{Q}_n^0)$ and g_n^0.

13.3.2 Parametric Submodel for Full-Data TMLE

Let $Q^0_{W,n}(\epsilon_1) = (1 + \epsilon_1 D^F_W(Q^0_n))Q^0_{W,n}$ be a parametric submodel through $Q^0_{W,n}$, and let

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

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

13.3.3 Obtaining a Targeted Estimate of Q_0

We define

$$\epsilon_n = \arg\min_{\epsilon} \sum_{i=1}^{n} q_0 L^F(Q^0_n(\epsilon))(W_{1i}, A_{1i}) + \frac{1 - q_0}{J} \sum_{j=1}^{J} L^F(1 - Q^0_n(\epsilon))(W^j_{2i}, A^j_{2i})$$

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

13.3.4 Estimator of the Target Parameter

Lastly, one evaluates the target parameter $\psi^*_n = \Psi^F(Q^*_n)$, where $Q^*_n = (\bar{Q}^1_n, Q^0_{W,n})$, by plugging $\bar{Q}^1_n$ and $Q^0_{W,n}$ into our substitution estimator to get the CCW-TMLE of ψ^F_0:

$$\psi^*_n = \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(q_0 \bar{Q}^1_n(1, W_{1,i}) + \frac{1 - q_0}{J} \sum_{j=1}^{J} \bar{Q}^1_n(1, W^j_{2,i}) \right) \right.$$
$$\left. - \left(q_0 \bar{Q}^1_n(0, W_{1,i}) + \frac{1 - q_0}{J} \sum_{j=1}^{J} \bar{Q}^1_n(0, W^j_{2,i}) \right) \right\}.$$

13.3.5 Calculating Standard Errors

Recall from Part I that the variance of our estimator is well approximated by the variance of the influence curve, divided by sample size n. Let IC^F be the influence curve of the full-data TMLE. We also showed that one can define IC^F as the full-data efficient influence curve given in (13.1). The case-control-weighted influence curve for the risk difference is then estimated by

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

Just as in Chap. 4, an estimate of the asymptotic variance of the standardized TMLE viewed as a random variable, using the estimate of the influence curve $IC_n(O)$, is given by $\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} IC_n^2(O_i)$.

13.4 Simulations

In the following simulation studies, we compare the CCW-TMLE to two other estimators to examine finite sample performance.

CCW-MLE. Case-control-weighted estimator of $\bar{Q}_0$ mapped to causal effect estimators by averaging over the case-control-weighted distribution of W. This is a case-control-weighted maximum likelihood substitution estimator of the g-formula (CCW-MLE) first discussed in van der Laan (2008a) and Rose and van der Laan (2008).

CCW-TMLE. The targeted case-control-weighted maximum likelihood substitution estimator of the g-formula discussed in the chapter.

IPTW estimator. Robins (1999a) and Mansson et al. (2007) discuss, under a rare disease assumption, the use of an "approximately correct" IPTW method for case-control study designs. It uses the estimated exposure mechanism among control subjects to update a logistic regression of Y on A. This estimator targets a nonparametrically nonidentifiable parameter, which indicates strong sensitivity to model misspecification for the exposure mechanism. Estimates of the risk difference and relative risk cannot be obtained using this method.

We limit our simulations in this chapter to the odds ratio since the IPTW estimator can only estimate this parameter.

Simulation 1. This first simulation study was based on a population of $N = 120{,}000$ individuals, where we simulated a one-dimensional covariate W, a binary exposure A, and an indicator Y. These variables were generated according to the following rules: $W \sim U(0, 1)$, $P_{X,0}(A \mid W) = \text{expit}(W^2 - 4W + 1)$, and $P_{X,0}(Y = 1 \mid A, W) = \text{expit}(1.2A - \sin W^2 + A \sin W^2 + 5A \log W + 5 \log W - 1)$. The resulting population had a prevalence probability of $q_0 = 0.035$, and exactly 4,165 cases. We sampled the

population using a varying number of cases and controls, and for each sample size we ran 1,000 simulations. The true value for the odds ratio was given by $OR = 2.60$.

For methods requiring an initial estimator of the conditional mean of Y, it was estimated using a correctly specified logistic regression and also a misspecified logistic regression with A and W as main terms. For methods requiring a fit for exposure mechanism, it was estimated using a correctly specified logistic regression and also a misspecified logistic regression with only the main term W.

Since we realistically generated A dependent on W, this led to substantial increases in efficiency in the targeted estimator when the initial estimator was misspecified and sample size grew, as it also adjusts for the exposure mechanism. This emphasizes the double robustness of the targeted estimators, and suggests that one should always target in practice. It is not surprising that when $\bar{Q}_n(A, W)$ was correctly specified, the relative efficiency of the targeted estimator (CCW-TMLE) was similar to its nontargeted counterpart (CCW-MLE). One should recall that correct specification in practice is unlikely and also note that this data structure is overly simplistic compared to real data. Even with this simple data structure, the IPTW estimators had the poorest overall efficiency. MSEs and relative efficiencies for the causal odds ratio are provided in Table 13.1.

When examining bias, it is clear that the IPTW estimators had the highest level of bias across all sample sizes, as observed in the bias plot displayed in Fig. 13.2. The CCW-MLE and CCW-TMLE with misspecified initial $\bar{Q}_n(A, W)$ had more bias than their correctly specified counterparts.

Table 13.1 Simulation results for the odds ratio. M is for misspecified $\bar{Q}_n(A, W)$ or $g_n(A \mid W)$ fit, C is for correctly specified $\bar{Q}_n(A, W)$ or $g_n(A \mid W)$. When two letters are noted in the "Fit" column, the first letter refers to $\bar{Q}_n(A, W)$ and the second to $g_n(A \mid W)$

Simulation 1	Fit	nC / nCo	250 / 250	500 / 500	500 / 1000	1000 / 1000	1000 / 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
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
CCW-TMLE RE	CC		1.27	3.62	14.58	8.40	32.03
	CM		1.26	3.62	14.57	8.40	31.97
	MC		1.96	4.63	16.68	9.52	31.91

Simulation 2	Fit	nC / nCo	100 / 250	250 / 250	250 / 500	500 / 500
IPTW MSE	M		404.40	3667.56	306.42	2433.62
IPTW RE	C		1.0	1.2	1.0	1.2
CCW-MLE RE	C		290	4200	570	5800
CCW-TMLE RE	CC		280	4100	570	5700
	CM		290	4100	570	5700

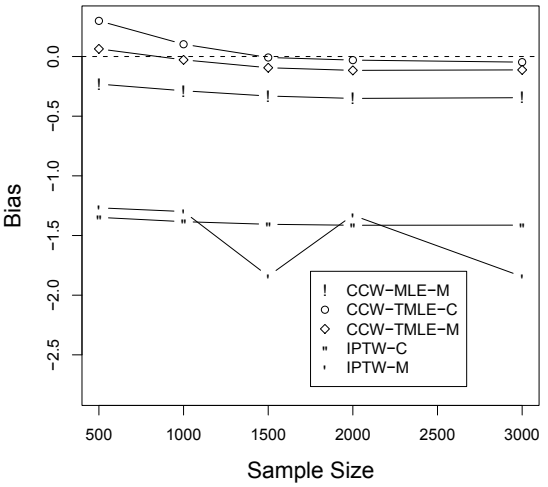

Fig. 13.2 Simulation 1 bias results. Bias results for the CCW-TMLE with misspecified $g_n(A \mid W)$ and the correctly specified CCW-MLE were excluded since values were the same as those for the TMLE with correctly specified $\bar{Q}_n(A, W)$ and $g_n(A \mid W)$.

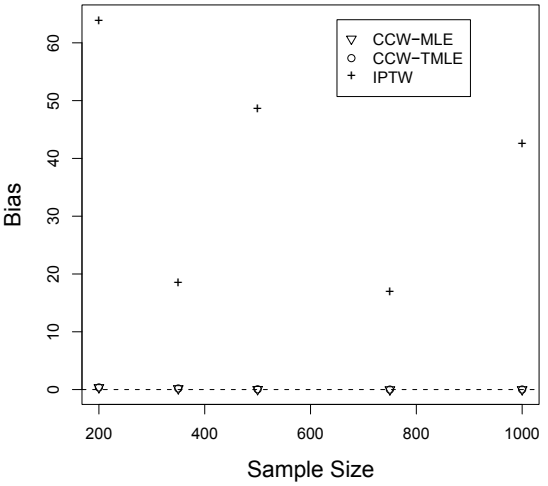

Fig. 13.3 Simulation 2 bias results. Bias results for the CCW-TMLE with misspecified $g_n(A \mid W)$ were excluded since those values were the same as those for the CCW-TMLE with correctly specified $\bar{Q}_n(A, W)$ and $g_n(A \mid W)$

Simulation 2. Our second set of simulations was based on a population of $N = 80{,}000$ individuals. The population had a binary exposure A, binary disease status Y, and a one-dimensional covariate W. These variables were generated according to the following rules: $W \sim U(0,1)$, $P_{X,0}(A \mid W) = \text{expit}(-5 \sin W)$, and $P_{X,0}(Y = 1 \mid A, W) = \text{expit}(2A - 25W + A \times W)$. The resulting population had a prevalence probability of $q_0 = 0.053$, exactly 4,206 cases. The true value for the odds ratio was given by $OR = 3.42$. The parameter was estimated using the same general methods as in the previous section, albeit with different fits for $\bar{Q}_n(A, W)$ and $g_n(A \mid W)$. The initial fit for each method requiring an estimate of $\bar{Q}_0(A, W)$ was estimated using a correctly specified logistic regression. For methods requiring a fit for exposure mechanism, it was estimated using a correctly specified logistic regression and also a misspecified logistic regression with W as a main term.

Results across the two case-control-weighted methods for the odds ratio were nearly identical, indicating again that when $\bar{Q}_n(A, W)$ 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 extremely inefficient in comparison. We theorized in van der Laan (2008a), and Mansson et al. (2007) demonstrated, that the IPTW procedure has a strong sensitivity to model misspecification. This result was observed in simulation 1, although the results in simulation 2 are more extreme. Results can be seen in Table 13.1 and Fig. 13.3.

13.5 Discussion

Case-control weighting provides a framework for the analysis of case-control study designs using TMLEs. We observed that the IPTW method was outperformed in conditions similar to a practical setting by CCW-TMLE in two simulation studies. The CCW-TMLE yields a fully robust and locally efficient estimator of causal parameters of interest. Model misspecification within this framework, with known or consistently estimated exposure mechanism, still results in unbiased and highly efficient CCW-TMLE. Further, in practice we recommend the use of super learner for the estimation of $\bar{Q}_0$. 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.

13.6 Notes and Further Reading

As previously discussed, conditional estimation of the odds ratio of disease given the exposure and baseline covariates is the prevalent method of analysis in case-control study designs. Key publications in the area of logistic regression in parametric statistical models for independent case-control study designs are Anderson (1972), Prentice and Pyke (1979), Breslow and Day (1980), and Breslow (1996). Green-

land (1981) and Holland and Rubin (1988) discuss another model-based method: the use of log-linear statistical models to estimate the marginal odds ratio. There are also multiple 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). We also refer the interested reader to Newman (2006) for a related IPTW-type method. This procedure builds on the standardization approach in order to weight exposed and unexposed controls using a regression of A on W.

Given the availability of city, state, and national databases for many diseases, including many cancers, knowledge of the prevalence probability is now increasingly realistic. The literature, going back to the 1950s, supports this. See, for example, Cornfield (1951, 1956). If the prevalence probability is not known, an estimate can be used in the CCW-TMLE, and this additional uncertainty can be incorporated into the standard errors. In situations where data on the population of interest may be sparse, the use of a range for the prevalence probability is also appropriate.

Other papers, in addition to Anderson (1972), discuss the use of $\log(q_0/(1-q_0))$ as an update to the intercept of a logistic regression, including Prentice and Breslow (1978), Greenland (1981), Morise et al. (1996), Wacholder (1996), and Greenland (2004). However, its use in practice remains limited. The adjustment is sometimes presented as a ratio of sampling fractions: $\log(P(\text{sampled} \mid Y = 1)/P(\text{sampled} \mid Y = 0))$, which reduces to $\log(q_0/(1-q_0))$.

This chapter was adapted from a previously published paper (Rose and van der Laan 2008). We refer readers to this paper for additional simulations where q_0 is estimated, and for a demonstration of the use of the influence curve for standard error estimation in a single simulated data set. We also refer readers to van der Laan (2008a) for the theoretical development of CCW-TMLE, as well as a formal discussion of the "approximately correct" IPTW estimator. The appendix of van der Laan (2008a) also discusses in detail the incorporation of the additional uncertainty from an estimated q_0 into the standard errors.

The complexity of a case-control study can vary. Additional designs include individually matched, incidence-density, and nested. Individually matched case-control studies are discussed in the next chapter, and prediction in nested case-control studies is discussed in Chap. 15. A TMLE for general two-stage designs, including so-called nested case-control designs, is presented in Rose and van der Laan (2011). Adaptations for incidence-density designs are discussed briefly in van der Laan (2008a) and will be further developed in future work.