Causal Inference for Case-Control Studies

By

Sherri Rose

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Mark van der Laan, Chair
Professor Nicholas Jewell
Professor Ira Tager

Spring 2011

Causal Inference for Case-Control Studies

Abstract

Causal Inference for Case-Control Studies

by

Sherri Rose

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Mark van der Laan, Chair

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. The data-generating experiment in case-control study designs involves an additional complexity called biased sampling. That is, one assumes the underlying experiment that randomly samples a unit from a target population, measures baseline characteristics, assigns an exposure, and measures a final binary outcome, but one samples from the conditional probability distribution, given the value of the binary outcome. One still desires to assess the causal effect of exposure on the binary outcome for the target population.

The targeted maximum likelihood estimator of a causal effect of treatment on the binary outcome based on such case-control studies is presented. Our proposed case-control-weighted targeted maximum likelihood estimator 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 targeted maximum likelihood estimator for a random sample into a method for case-control sampling.

Individually matched case-control study designs are commonly implemented in the field of public health. While matching is intended to eliminate confounding, the main *potential* benefit of matching in case-control studies is a gain in efficiency. We investigate the use of the case-control-weighted targeted maximum likelihood estimator to estimate causal effects in matched case-control study designs. We also compare the case-control-weighted targeted maximum likelihood estimator in matched and unmatched designs in an effort to determine which design yields the most information about the causal effect. In many practical situations where a causal effect is the parameter of interest, researchers may be better served using an unmatched design.

We also consider two-stage sampling designs, including so-called nested case-control studies, where one takes a random sample from a target population and completes measurements on each subject in the first stage. The second stage involves drawing a subsample from the original sample, collecting additional data on the subsample. This data structure can be viewed as a missing data structure on the full-data structure collected in the second stage of the study. We propose an inverse-probability-of-censoring-weighted targeted maximum likelihood estimator in two-stage sampling designs. Two-stage designs are also common for prediction research questions. We present an analysis using super learner in nested case-control data from a large Kaiser Permanente database to generate a function for mortality risk prediction.

Contents

Acknowledgements

I have had an amazing experience at Berkeley, including many opportunities not typical for a graduate student. I must start by thanking my advisor and dissertation chair **Mark J. van der Laan**. I've learned so much working with Mark over the past few years, and he's opened my mind to new ways of thinking. I didn't think it was possible, but he's made me even more excited about the field of statistics. I also want to thank my dissertation committee members **Nicholas P. Jewell** and **Ira B. Tager**. Their insights were essential to this research.

I thank the following colleagues and mentors for their ultimate impact on my graduate experience: **Antoine Chambaz** (Université Paris Descartes), **Lori Chibnik** (Boston University SIBS 2004), **Sandrine Dudoit** (UC Berkeley), **Justin Fisher** (GWU), **Donald Garecht** (Marlton Middle School: 1996–97), **Alan E. Hubbard** (UC Berkeley), **Terry Hufford** (GWU), **Subrata Kundu** (GWU), **Maya L. Petersen** (UC Berkeley), **Steve Selvin** (UC Berkeley), **Lorraine Stern** (Case Western), and **Lisa Sullivan** (Boston University SIBS 2004).

I need to thank **Burke J. Bundy** (♡) for being an incredibly supportive partner. Loving an academic is no small chore and he is relentless in his devotion to my dreams. My PhD involved many late nights, busy mornings, and an all-consuming focus. Burke was there for all of it, offering humor and compassion. I would also like to thank the following friends and family: **Paula** and **Andy Clearfield**, **Harold** and **Judy Diamond**, **Susanne M. Maurer** (§), **Jennifer L. Poirier**, **David** and **Amy Weinstein**, **Norman** and **Mona Weinstein**, **Henry** and **Beth White**, and **Adrienne Wilson**. For our Trooper 🐶.

Chapter 1

Introduction: Case-Control Studies

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., the probability of having the disease ≈ 0), 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. However, 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 Y given the exposure of interest A and measured covariates W.

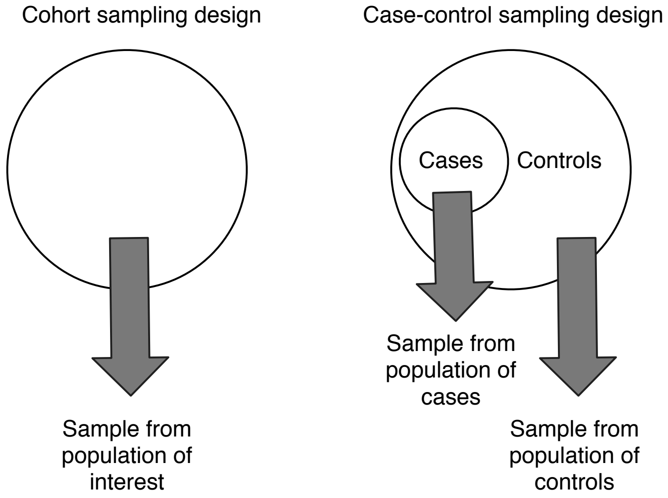

Cohort sampling design
Case-control sampling design
Cases
Controls
Sample from population of cases
Sample from population of interest
Sample from population of controls

Figure 1.1: Case-control sampling design

1.1 Independent Designs

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). Greenland (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).

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 among control subjects as a weight to update a logistic regression of disease status on exposure. This inverse-probability-of-treatment-weighted (IPTW) estimator does not require the knowledge of prevalence probability, only that the prevalence probability is close to zero. The IPTW estimator presented by Robins (1999) targets a nonparametrically nonidentifiable parameter, which indicates strong sensitivity towards model misspecification for the exposure mechanism. Additionally, the causal effect estimates of the risk difference and relative risk cannot be obtained using this method. This IPTW estimator will be examined further in Chapter 3. 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.

The case-control-weighted targeted maximum likelihood estimator (CCW-TMLE) (van der Laan 2008a; Rose and van der Laan 2008, 2009; van der Laan and Rose 2011) we propose relies on knowledge of the true prevalence probability $P_{X,0}(Y = 1) \equiv q_0$, 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 targeted maximum likelihood estimator (TMLE) for a random sample (van der Laan and Rubin 2006; van der Laan and Rose 2011) into a method for case-control sampling. The CCW-TMLE, presented in Chapter 3, 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. Additional papers on TMLE in observational and experimental studies include Bembom and van der Laan (2007), Moore and van der Laan (2009a,b,c), Bembom et al. (2009), Polley and van der Laan (2009), Rosenblum et al. (2009), van der Laan and Gruber (2010), Gruber and van der Laan (2010a,b), Rosenblum and van der Laan (2010), Wang et al. (2010), Stitelman and van der Laan (2010, 2011a,b), and Rose and van der Laan (2011).

We also propose the use of super (machine) learning methods (van der Laan et al. 2007) within the CCW-TMLE procedure. The super learner allows researchers to use multiple algorithms to outperform a single algorithm in realistic nonparametric and semiparametric statistical models based on actual knowledge. It's use within CCW-TMLE will be described in Chapter 3. Super learner is a generalization of the stacking algorithm introduced in the neural networks context by Wolpert (1992) and adapted to the regression context by Breiman (1996), and its name was introduced due to the theoretical oracle property and its consequences as presented in van der Laan and Dudoit (2003). The stacking algorithm is examined in LeBlanc and Tibshirani (1996) and the relationship to the model-mix algorithm of Stone (1974) and the predictive sample-reuse method of Geisser (1975) is discussed. Recent literature on aggregation and ensemble learners includes Tsybakov (2003), Juditsky et al. (2005), Bunea et al. (2006, 2007a,b), and Dalalyan and Tsybakov (2007, 2008). Targeted Learning, including super learner and the TMLE, while be presented in Chapter 2.

The population under study should be clearly defined. As such, the prevalence probability is then truly basic information about a population of interest. Given the availability of city, state, and national databases for many diseases, 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 (van der Laan 2008a; Rose and van der Laan 2008). In situations where data on the population of interest may be sparse, the use of a range for the prevalence probability is also appropriate.

The use of the prevalence probability to eliminate the bias of case-control sampling design has previously been discussed as update to a logistic regression model with the intercept update $\log(q_0/(1-q_0))$ (Anderson 1972; Prentice and Breslow 1978; Greenland 1981; Morise et al. 1996; Wacholder 1996; Greenland 2004). However, its use in practice remains limited. The intercept-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 $\log(q_0/(1-q_0))$.

The complexity of a case-control study can vary. Additional designs include individually matched, incidence-density, and nested. A literature review for individually matched case-control studies follows in the next section, and a CCW-TMLE for individually matched studies is presented in Chapter 4, which was adapted from Rose and van der Laan (2009). A TMLE for general two-stage designs, including so-called nested case-control designs, is presented in Chapter 5, which was adapted from Rose and van der Laan (2011). A literature review for two-stage designs is presented in Section 1.3. Chapter 3 was adapted from Rose and van der Laan (2008). Methdology for incidence-density designs is discussed briefly in van der Laan (2008a) and are an area of future work.

1.2 Individually Matched Designs

In an individually matched case-control study, the population of interest is identified, and cases are randomly sampled or selected based on particular inclusion criteria. Each of these cases is then matched to one or more controls based on a variable (or variables) *believed* to be a confounder. There is a large collection of literature devoted to the topic of individual matching in case-control study designs, and discussion of the advantages and disadvantages of matching goes back more than 40 years. While some literature cites the purpose of matching as improving validity, later publications (Kupper et al. 1981; Rothman and Greenland 1998) demonstrate that matching has a greater impact on efficiency over validity. Costanza (1995) notes that matching on confounders in case-control studies does nothing to remove the confounding. Similarly, Rothman and Greenland (1998) discuss that matching cannot control confounding in case-control study designs but can, in fact, introduce bias. Methodologists in the literature stress that it is often possible and preferred for confounders to be *adjusted for* in the analysis instead of matching in case-control designs (Schlesselman 1982; Vandenbrouke et al. 2007).

Matching has a substantial impact on the study sample; most notably, it creates a sample of controls that is not representative of exposure in the population or the population as a whole. The effect of the matching variable can no longer be studied directly, and the exposure frequency in the control sample will be shifted towards that of the cases (Rothman and Greenland 1998). The matched sampling leads to a balanced number of cases and controls across the levels of the selected matching variables. This balance can reduce the variance in the parameter of interest, which improves statistical efficiency. A study with a randomly selected control group may yield some strata with an imbalance of cases and controls.

It is important to add, however, that matching in case-control studies can lead to gains *or* losses in efficiency (Kupper et al. 1981; Rothman and Greenland 1998). Matching variables are chosen a priori on the belief that they confound the relationship between exposure and disease. If controls are matched to cases based on a variable that is not a true confounder, this can impact efficiency. For example, if the matching variable is associated not with disease but with the exposure, this will increase the variance of the estimator compared to an unmatched design. Here, the matching leads to larger numbers of exposure-concordant case-control pairs, which are not informative in the analysis, leading to an increase in variance. If the matching variable is only associated with disease, there is often a loss of efficiency as well (Schlesselman 1982). If the matching variable is along the causal pathway between disease and exposure, then matching will contribute bias that cannot be removed in the analysis (Vandenbrouke et al. 2007).

The number of matching variables should also be reduced to as few as possible. As the number of matching variables grows, the cases and controls will become increasingly similar with respect to the exposure of interest, and the study may produce a spurious result or provide no information (Breslow and Day 1980). Additionally, when matching on more than one variable, matching variables should not be strongly correlated with each other (Schlesselman 1982).

Cochran (1953) demonstrates the efficiency of matched designs. However, as noted by McKinlay (1977), Cochran's result can be misleading. Comparisons between matched and unmatched study designs are often made with *equal* sample sizes and no other method of covariate adjustment. In a matched design, controls may be discarded if they do not match a particular case on the variable or variables of interest. Multiple controls may be discarded per case, depending on the variables of interest (Freedman 1950; Cochran 1965; McKinlay 1977). In many cases, if the discarded controls were available to be rejected in the matched study, they would be available for an unmatched design in the same investigation (Billewicz 1965; McKinlay 1977). Therefore, it is often more appropriate to compare the efficiencies of matched case-control studies of size n to randomly selected case-control studies of size n+*number of discarded controls*.

Kupper et al. (1981) performed a variety of simulations to demonstrate the impact of matching on efficiency. They found that in situations where confounding was present, the confidence intervals for matched studies were smaller than unmatched studies unless the odds ratio and the exposure of interest were large. However, the confidence intervals for the samples with randomly selected controls were always shorter when the number of controls was at least twice that of the cases. This is an important result, as efficiency is often touted as the benefit of an individually matched case-control study design, and the discussion above that comparisons between matched case-control studies of size n should be made to randomly selected case-control studies of size n+*number of discarded controls*.

The predominant method of analysis in individually matched case-control studies is conditional logistic regression in a parametric statistical model. The logistic regression model for matched case-control studies differs from unmatched studies in that it allows the intercept to vary among the matched units of cases and controls. The matching variable is not included in the model (Breslow et al. 1978; Holford et al. 1978; Breslow and Day 1980; Schlesselman 1982). In order to estimate an effect of exposure A with conditional logistic regression, the case and control must be discordant on A. Additionally, if information for a variable is missing for a case (or control), the corresponding control (or case) information is discarded (Breslow and Day 1980; Schlesselman 1982). Rothman and Greenland (1998) and Greenland (2004) demonstrate the use of standardization in case-control studies, which estimate marginal effects with population or person-time averaging.

Gefeller et al. (1998) performed a literature review of case-control studies published between 1955 and 1994 in three main epidemiology journals: *American Journal of Epidemiology, International Journal of Epidemiology, and the Journal of Epidemiology and Community Health.* They found that, among these journals, there was a decreasing trend in the percentage of individually matched case-control studies published (71.7% in the years preceding 1981, 65.5% in 1985, 46.9% in 1989, and 46.4% in 1994), and an increasing percentage of frequency matched studies (5.0% in the years preceding 1981, 9.1% in 1985, 16.3% in 1989, and 26.2% in 1994). Interestingly, the percentage of case-control studies using no matching stayed relatively constant with no obvious trend (averaging 29.3%, and ranging from 23.2% to 36.7%). Unfortunately, they found substantial evidence that individually matched studies were being performed without the appropriate matched analysis: only 74% of studies from 1994 used conditional logistic regression if logistic regression was the chosen method of analysis. A later analysis of medical literature in Medline, Rahman (2003), indicated that 5.3% of individually matched case-control studies used an unconditional logistic regression for those selecting logistic regression models. The review in Gefeller et al. (1998) indicates that unmatched case-control studies, at least in epidemiology, are in the minority. This should be questioned given the overwhelming agreement in the literature that matching is not frequently justified for case-control study designs.

The consensus in the literature indicates that there are very few circumstances where individual matching is indeed warranted. Case-control studies with a very small number of cases may benefit from individual matching, as a randomly selected control group from even a well-defined population of interest may be uninformative on many variables of interest (Schlesselman 1982; Costanza 1995). Matching is cited as necessary by many authors when the investigators expect the distribution of the matching variable to differ drastically between the cases and the controls. It may be this reason that draws many investigators towards a matched design, perhaps without appropriate consideration of the disadvantages or definition of the population of interest. Are the direct time and labor costs of individually matching controls worth the potential gain in efficiency? Will the potential efficiency gain outweigh a delay of x months or years in the collection of study data? Is a gain in efficiency truly likely? Might obtaining a larger, randomly sampled group of controls be sufficient to ensure coverage over the distribution of the confounder? It is therefore important to continue to disseminate the implications of individually matched case-control study designs to researchers, as Rothman and Greenland (1998) note that *"people match on a variable (e.g. sex) simply because it is the 'expected thing to do' and they might lose credibility for not matching."* When researchers make design and analysis decisions based on these types of considerations, their research may suffer.

1.3 Two-Stage Designs

Two-stage designs, including nested case-control studies, have been discussed and developed in previous literature, including Neyman (1938), Cochran (1963), Mantel (1973), Kupper et al. (1975), Liddell et al. (1977), Thomas (1977), and Breslow et al. (1983). Advantages can include reduction in costs associated with collecting data on the entire cohort and minimal losses in efficiency (Ernster 1994; Rothman and Greenland 1998; Essebag et al. 2003; Hak et al. 2004; Vittinghoff and Bauer 2006).

1.3.1 Effect estimation

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

A recent paper by Wang et al. (2009) presents causal effect estimators using estimating equation methodology where the outcome Y, exposure A, and a subset S of covariates W are measured in the first stage (V includes Y, A, and S). They consider the same two-stage design, where one measures $V = (S, Y, A)$ on everyone in the sample, and $X = (S, Y, A, W)$ on the subjects in the validation sample defined by $\Delta = 1$, where the missingness mechanism is known. The Wang et al. article focuses on estimation of EY_a under the consistency assumption $Y = Y_A$, the randomization assumption, A is independent of Y_a, given (W, S), so that $EY_a = E_{S,W} E_{X,0}(Y \mid A = a, S, W)$, and a parametric model for the treatment mechanism $\Pi(S, W) = P(A = 1 \mid S, W)$. We refer the interested reader to the appendix of Chapter 5 for a detailed discussion of the relationships between the estimators presented in Wang et al. (2009) and the TMLE we will present.

1.3.2 Prediction

Prediction has been used most notably to generate tables for risk of heart disease (Kannel et al. 1976; Anderson et al. 1991; Ramsay et al. 1995, 1996; Wilson et al. 1998; Jackson 2000) and breast cancer (Gail et al. 1989; Costantino et al. 1999; Tyrer et al. 2004; Barlow et al. 2006). An existing method for prediction in parametric statistical models with nested case-control samples is intercept adjustment discussed in Section 1.1. The addition of $\log(P_{X,0}(\Delta = 1 \mid Y = 1)/P_{X,0}(\Delta = 1 \mid Y = 0))$, or equivalently $\log(q_0/(1 - q_0))$, to the intercept in a logistic regression yields the true logistic regression function $P_{X,0}(Y = 1 \mid W)$, assuming the statistical model is correctly specified. Here Δ denotes the indicator of inclusion in the nested case-control sample, and the value q_0 is the prevalence probability $P_{X,0}(Y = 1) = q_0$ (Anderson 1972; Prentice and Breslow 1978; Greenland 1981; Wacholder 1996; Morise et al. 1996; Greenland 2004). We will use the super learner (van der Laan et al. 2007), also discussed in Section 1.1, to provide a more flexible method for prediction in two-stage nested case-control data.

1.4 Conclusion

There had been relatively little work completed in the area of causal inference and prediction for case-control study designs. Given the popularity of these designs in the public health and medical literature, the advances targeted learning using super learner and TMLE offer, in conjunction with case-control weighting, are substantial.

Chapter 2

Road Map for Targeted Learning

One of the great open problems across many fields of research has been obtaining causal effects from data. Data are typically sampled from a population of interest since collecting data on the entire population is not feasible. Frequently, the researcher is not interested in merely studying association or correlation in this sample data; she wants to know whether an exposure (or treatment) causes the outcome in the population of interest. If one can show that the exposure causes the outcome, we can then impact the outcome by intervening on the exposure.

The often quoted "ideal experiment" is one that cannot be conducted in real life. Let us say we are interested in studying the causal effect of a toxin on death from cancer within 5 years. In an ideal experiment, we intervene and set the exposure to *exposed* for each subject. All subjects are followed for 5 years, where the outcome under this exposure is recorded. We then go back in time to the beginning of the study, intervene, and set all subjects to *not exposed* and *follow them under identical conditions* until the end of the study, recording the outcome under this exposure.

Let's assume in principle there is a system where this ideal experiment could have been conducted. This experiment generates random variables. Say the experiment is that we sample a subject (i.e., draw a random variable) from a population and take several measurements on this subject. This experiment is repeated multiple times until we have sampled an a priori specified number (representing the sample size) of subjects. These random variables also have a true underlying probability distribution. Our observed data are realizations of these random variables. If we were to conduct our repeated experiment again, we would observe different realizations of these random variables.

Any knowledge we have about how these observed data were generated is referred to as a model. For example, it might be known that the data consist of observations on a number of independent and identically distributed (i.i.d.) random variables. So, our data are i.i.d. random variables, but the probability distribution of the random variable is typically completely unknown. This is also information we incorporate into our model. We refer to this as a nonparametric model for the probability distribution of the random variable. (Do note, however, that assuming the data vector is i.i.d. in our nonparametric model is a real assumption, although one we will always make in this dissertation.) Our model should always reflect true knowledge about the probability distribution of the data, which may often be a nonparametric model, or a semiparametric model that makes some additional assumptions. For example, perhaps it is known that the probability of death is monotonically increasing in the levels of exposure, and we want to include this information in our model.

The knowledge we have discussed thus far regarding our model pertains to our observed data and what we call the statistical model. The statistical model is, formally, the collection of possible probability distributions. The model may also contain extra information in addition to the knowledge contained in the statistical model. We want to relate our observed data to a causal model. We can do this with additional assumptions, and we refer to a statistical model augmented with these additional causal assumptions as the model for the observed data. These additional assumptions allow us to define the system where this ideal experiment could have been conducted. We can describe the generation of the data with nonparametric structural equations, intervene on exposure and set those values to *exposed* and *not exposed*, and then see what the (counterfactual) outcomes would have been under both exposures. This underlying causal model allows us to define a causal effect of treatment or exposure.

One now needs to specify the relation between the observed data on a unit and the full data generated in the causal model. For example, one might assume that the observed data corresponds with observing all the variables generated by the system of structural equations that make up the causal model, up to background factors that enter as error terms in the underlying structural equations. The specification of the relation between the observed data and this underlying causal model allows one now to assess if the causal effect of interest can be identified from the probability distribution of the observed data. If that is not possible, then we state that the desired causal effect is not identifiable. If, on the other hand, our causal assumptions allow us to write the causal effect as a particular feature of the probability distribution of the observed data, then we have identified a target parameter of the probability distribution of the observed data that can be interpreted as a causal effect.

Let's assume that the causal effect is identifiable from the observed data. Our parameter of interest, here the causal effect of a toxin on death from cancer within 5 years, is now a parameter of our true probability distribution of the observed data. This definition as a parameter of the probability distribution of the observed data does not rely on the causal assumptions coded by the underlying causal model describing the ideal experiment for generating the desired full data, and the link between the observed data and the full data. Thus, if we ultimately do not believe these causal assumptions, the parameter is still an interesting statistical parameter. Our next goal becomes estimating this parameter of interest.

The open problem addressed in this dissertation is the estimation of interesting parameters of the probability distribution of the data in case-control study designs. This need not only be (causal) effect measures. Another problem researchers are frequently faced with is the generation of functions for the prediction of rare outcomes.

For these problems, we do not make causal assumptions, but still define our realistic nonparametric or semiparametric statistical model based on actual knowledge. We view effect and prediction parameters of interest as features of the probability distribution of our data, well defined for each probability distribution in the nonparametric or semiparametric model. Statistical learning from data is concerned with efficient and unbiased estimation of these features and with an assessment of uncertainty of the estimator.

In Chapters 3–5 we develop targeted learning for case-control studies. In this chapter, we will develop the following concepts, as part of the general road map for targeted learning in observational and experimental data:

Defining the data, model, and parameter. We will define our random variable, which we observe n times, and corresponding probability distribution of interest. By defining a structural causal model (SCM), we specify a model for underlying counterfactual outcome data, representing the data one would be able to generate in an ideal experiment. This is a translation of our knowledge about the data-generating process into causal assumptions. We can define our target parameter in our SCM, i.e., as a so-called causal effect of an intervention on a variable A on an outcome Y. The SCM also generates the observed data O, and one needs to determine if the target parameter can be identified from the distribution P_0 of O alone. In particular, one needs to determine what additional assumptions are needed in order to obtain such identifiability of the causal effect from the observed data.

Super learning for prediction. The first step in our estimation procedure is an initial estimate for the part of the data-generating distribution P_0 required to evaluate the target parameter. This estimator needs to recognize that P_0 is only known to be an element of a semiparametric statistical model. That is, we need estimators that are able to truly learn from data, allowing for flexible fits with increased amounts of information. Super learning provides an optimal approach to estimation of P_0 (or infinite-dimensional parameters thereof) in semiparametric statistical models. Since prediction can be a research question of interest in itself, super learning for prediction is useful as a standalone tool as well.

TMLE. With an initial estimate of the relevant part of the data-generating distribution obtained using super learning, we present the remainder of the TMLE. The second stage of TMLE updates this initial fit in a step targeted towards making an optimal bias–variance tradeoff for the parameter of interest, instead of the overall probability distribution P_0. This results in a targeted estimator of the relevant part of P_0, and thereby in a corresponding substitution estimator of $\Psi(P_0)$.

2.1 Data, Model, Parameter

2.1.1 Data

The data are n i.i.d. observations of a random variable $O \sim P_0$, where O has probability distribution P_0. For a simple example, suppose our data structure is $O = (W, A, Y) \sim P_0$. We have a covariate or vector of covariates W, an exposure A, and an outcome Y. The random variables $O_1, \ldots, O_n$ might be the result of randomly sampling n subjects from a population of patients, collecting baseline characteristics W, assigning exposure A, and following the subjects and measuring outcome Y. The case-control data structure for independent case-control study designs is discussed in Chapter 3, individually matched case-control study designs in Chapter 4, and nested two-stage designs in Chapter 5.

2.1.2 Model

We are considering the general case that one observed n i.i.d. copies of a random variable O with probability distribution P_0. The data-generating distribution P_0 is also known to be an element of a statistical model $\mathcal{M}$, which we write $P_0 \in \mathcal{M}$. A statistical model $\mathcal{M}$ is the set of possible probability distributions for P_0; it is a collection of probability distributions. In this dissertation, the distribution of our data is simply known to be an element of a nonparametric statistical model $\mathcal{M}$.

We specify a set of endogenous variables $X = (X_j : j)$. In a very simple example, we might have $j = 1, \ldots, J$, where $J = 3$. Thus, $X = (X_1, X_2, X_3)$. We can rewrite X as $X = (W, A, Y)$ if we say $X_1 = W$, $X_2 = A$, and $X_3 = Y$. For each endogenous variable X_j one specifies the parents of X_j among X, denoted $Pa(X_j)$. We denote a collection of exogenous variables by $U = (U_{X_j} : j)$. One assumes that X_j is some function of $Pa(X_j)$ and an exogenous U_{X_j}:

$$X_j = f_{X_j}(Pa(X_j), U_{X_j}),\ j = 1 \ldots, J.$$

The collection of functions f_{X_j} indexed by all the endogenous variables is represented by $f = (f_{X_j} : j)$. Together with the joint distribution of U, these functions f_{X_j}, specify the data-generating distribution of (U, X) as they describe a deterministic system of structural equations (one for each endogenous variable X_j) that deterministically maps a realization of U into a realization of X. In an SCM one also refers to some of the endogenous variables as intervention variables. The SCM assumes that intervening on one of the intervention variables by setting their value, thereby making the function for that variable obsolete, does not change the form of the other functions. The functions f_{X_j} are often unspecified, but in some cases it might be reasonable to assume that these functions have to fall in a certain more restrictive class of functions. Similarly, there might be some knowledge about the joint distribution of U. The set of possible data-generating distributions of (U, X) can be obtained by varying the structural equations f over all allowed forms, and the distribution of the errors U over all possible error distributions defines the SCM for the full-data (U, X), i.e., the SCM is a statistical model for the random variable (U, X).

The corresponding SCM for the observed data O also includes specifying the relation between the random variable (U, X) and the observed data O, so that the SCM for the full data implies a parameterization of the probability distribution of O in terms of f and the distribution P_U of U. This SCM for the observed data also implies a statistical model for the probability distribution of O.We have the functions $f = (f_W, f_A, f_Y)$ and the exogenous variables $U = (U_W, U_A, U_Y)$. The values of W, A, and Y are deterministically assigned by U corresponding to the functions f. We could specify our structural equation models, based on investigator knowledge, as

$$\begin{aligned} W &= f_W(U_W), \\ A &= f_A(W, U_A), \\ Y &= f_Y(W, A, U_Y), \end{aligned} \tag{2.1}$$

where no assumptions are made about the true shape of f_W, f_A, and f_Y. These functions f are nonparametric.

We may assume that U_A is independent of U_Y, given W, which corresponds with believing that there are no unmeasured factors that predict both A and the outcome Y: this is often called the no unmeasured confounders assumption (discussed later in this chapter). This SCM represents a semiparametric statistical model for the probability distribution of the errors U and endogenous variables $X = (W, A, Y)$. We assume that the observed data structure $O = (W, A, Y)$ is actually a realization of the endogenous variables (W, A, Y) generated by this system of structural equations. This now defines the SCM for the observed data O. It is easily seen that any probability distribution of O can be obtained by selecting a particular data-generating distribution of (U, X) in this SCM. Thus, the statistical model for P_0 implied by this SCM is a nonparametric model. As a consequence, one cannot determine from observing O if the assumptions in the SCM contradict the data. One states that the SCM represents a set of nontestable causal assumptions we have made about how the data were generated in nature.

We can draw a causal graph from our SCM, which is a visual way to describe some of the assumptions made by the model and the restrictions placed on the joint distribution of the data (U, X). Causal graphs cannot encode every assumption we make in our SCM, and, in particular, the identifiability assumptions derived from causal graphs alone are not specific for the causal parameter of interest. Identifiability assumptions derived from a causal graph will thus typically be stronger than required. In addition, the link between the observed data and the full-data model represented by the causal graph is often different than simply stating that O corresponds with observing a subset of all the nodes in the causal graph. In this case, the causal graph itself cannot be used to assess the identifiability of a desired causal

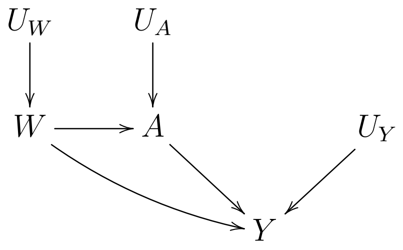


Figure 2.1: A possible causal graph for (2.1)

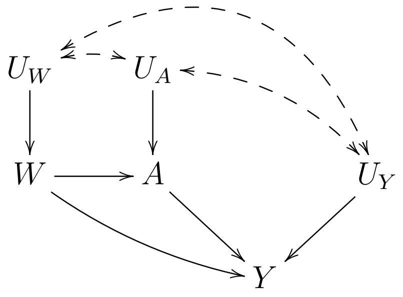


Figure 2.2: A causal graph for (2.1) with no assumptions on the distribution of P_U

parameter from the observed data distribution.

We previously mentioned the typically nontestable causal assumptions made by an SCM. We make the first assumption by defining the parents $Pa(X_j)$ for each endogenous X_j. The second is any set of assumptions about the joint distribution P_U of the exogenous variables. The assumptions made based on actual knowledge concerning the relationships between variables [i.e., defining the parents $Pa(X_j)$ for each endogenous X_j] are displayed in our causal graph through the presence and absence of directed arrows. In Fig. 2.1, the direction of the arrows is defined by the assignment of the parents to each node, including the time ordering assumed during the specification of (2.1). The assumptions on the distribution P_U are reflected in causal graphs through dashed double-headed arrows between the variables U. In Fig. 2.1 there are no arrows between the $U = (U_W, U_A, U_Y)$. Therefore, (2.1) included the assumption of joint independence of the endogenous variables U, which is graphically displayed by the lack of arrows. This is not an assumption one is usually able to make based on actual knowledge. More likely, we are able to make few or no assumptions about the distribution of P_U. For (2.1), with no assumptions about the distribution of P_U, our causal graph would appear as in Fig. 2.2.

2.1.3 Parameter

The estimation problem requires the description of a target parameter of P_0 one wishes to learn from the data. This definition of a target parameter requires specification of a mapping Ψ one can then apply to P_0. This mapping Ψ needs to be defined on any possible probability distribution in the statistical model $\mathcal{M}$. Thus Ψ maps any $P \in \mathcal{M}$ into a vector of numbers $\Psi(P)$. We write the mapping as $\Psi : \mathcal{M} \to \mathbb{R}^d$ for a d-dimensional parameter. We introduce ψ_0 as the evaluation of $\Psi(P_0)$, i.e., the true value of our parameter. The statistical estimation problem is now to map the observed data $O_1, \ldots, O_n$ into an estimator of $\Psi(P_0)$ that incorporates the knowledge that $P_0 \in \mathcal{M}$, accompanied by an assessment of the uncertainty in the estimator. For example, with the experimental unit-specific data structure $O = (W, A, Y) \sim P_0$, the risk difference is the following function of the distribution P_0 of O:

$$\Psi(P_0) = E_{W,0}[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W)],$$

where $E_0(Y \mid A = a, W)$ is the conditional mean of Y given $A = a$ and W, with binary A.

We can define a causal target parameter of interest as a parameter of the distribution of the full-data (U, X) in the SCM. Formally, we denote the SCM for the full-data (U, X) by $\mathcal{M}^F$, a collection of possible $P_{U,X}$ as described by the SCM. In other words, $\mathcal{M}^F$, a model for the full data, is a collection of possible distributions for the underlying data (U, X). Ψ^F is a mapping applied to a $P_{U,X}$ giving $\Psi^F(P_{U,X})$ as the target parameter of $P_{U,X}$. This mapping needs to be defined for each $P_{U,X}$ that is a possible distribution of (U, X), given our assumptions coded by the posed SCM. In this way, we state $\Psi^F : \mathcal{M}^F \to \mathbb{R}^d$, where $\mathbb{R}^d$ indicates that our parameter is a vector of d real numbers. The SCM $\mathcal{M}^F$ consists of the distributions indexed by the deterministic function $f = (f_{X_j} : j)$ and distribution P_U of U, where f and this joint distribution P_U are identifiable from the distribution of the full-data (U, X). Thus the target parameter can also be represented as a function of f and the joint distribution of U.

Recall our example with data structure $O = (W, A, Y)$ and SCM given in (2.1) with no assumptions about the distribution P_U. We can define $Y_a = f_Y(W, a, U_Y)$ as a random variable corresponding with intervention $A = a$ in the SCM. The marginal probability distribution of Y_a is thus given by

$$P_{U,X}(Y_a = y) = P_{U,X}(f_Y(W, a, U_Y) = y).$$

The causal effect of interest for a binary A (suppose it is the causal risk difference) could then be defined as a parameter of the distribution of (U, X) given by

$$\Psi^F(P_{U,X}) = E_{U,X}Y_1 - E_{U,X}Y_0.$$

In other words, $\Psi^F(P_{U,X})$ is the difference of marginal means of counterfactuals Y_1 and Y_0.

We will define our causal target parameter as a parameter of the distribution of the data (U, X) under an intervention on one or more of the structural equations in f. The intervention defines a random variable that is a function of (U, X), so that the target parameter is $\Psi^F(P_{U,X})$. We discussed the “ideal experiment” which we cannot conduct in practice, where we observe each subject's outcome at all levels of A under identical conditions. Intervening on the system defined by our SCM describes the data that would be generated from the system at the different levels of our intervention variable (or variables). For example, in our simple example, we can intervene on the exposure A in order to observe the results of this intervention on the system. By assumption, intervening and changing the functions f_{X_j} of the intervention variables does not change the other functions in f. With the SCM given in (2.1) we can intervene on f_A and set $a = 1$:

$$\begin{aligned} W &= f_W(U_W), \\ a &= 1, \\ Y_1 &= f_Y(W, 1, U_Y). \end{aligned}$$

We can also intervene and set $a = 0$:

$$\begin{aligned} W &= f_W(U_W), \\ a &= 0, \\ Y_0 &= f_Y(W, 0, U_Y). \end{aligned}$$

The intervention defines a random variable that is a function of (U, X), namely, $Y_a = Y_a(U)$ for $a = 1$ and $a = 0$. The probability distribution of the (X, U) under an intervention is called the postintervention distribution. Our target parameter is a parameter of the postintervention distribution of Y_0 and Y_1, i.e., it is a function of these two postintervention distributions, namely, some difference. Thus, the SCM for the full data allows us to define the random variable $Y_a = f_Y(W, a, U_Y)$ for each a, where Y_a represents the outcome that would have been observed under this system for a particular subject under exposure a. Thus, with the SCM we can carry out the "ideal experiment" and define parameters of the distribution of the data generated in this perfect experiment, even though our observed data are only the random variables $O_1, \ldots, O_n$.

We would ideally like to see each individual's outcome at all possible levels of exposure A. The study is only capable of collecting Y under one exposure, the exposure the subject experiences. For our binary exposure in the example above, we have $(Y_a : a)$, with $a \in \mathcal{A}$, and where $\mathcal{A}$ is the set of possible values for our exposure. Here, this set is simply $\{0, 1\}$, but in other examples it could be continuous or otherwise more complex. Thus, in our example, for each realization u, which might correspond with an individual randomly drawn from some target population, by intervening on (2.1), we can generate so-called counterfactual outcomes $Y_1(u)$ and $Y_0(u)$. These counterfactual outcomes are implied by our SCM; they are consequences of it. That is, $Y_0(u) = f_Y(W, 0, u_Y)$, and $Y_1(u) = f_Y(W, 1, u_Y)$, where $W = f_W(u_W)$ is also implied by u. The random counterfactuals $Y_0 = Y_0(U)$ and $Y_1 = Y_1(U)$ are random through the probability distribution of U. Our target parameter is a function of the probability distributions of these counterfactuals: $E_0Y_1 - E_0Y_0$.

Establishing identifiability. We want to be able to write $\Psi^F(P_{U,X,0})$ as $\Psi(P_0)$ for some parameter mapping Ψ. Since the true probability distribution of (U, X) can be any element in the SCM $\mathcal{M}^F$, and each such choice $P_{U,X}$ implies a probability distribution $P(P_{U,X})$ of O, this requires that we show that $\Psi^F(P_{U,X}) = \Psi(P(P_{U,X}))$ for all $P_{U,X} \in \mathcal{M}^F$. This step involves establishing possible additional assumptions on the distribution of U, or sometimes also on the deterministic functions f, so that we can identify the target parameter from the observed data distribution. Thus, for each probability distribution of the underlying data (U, X) satisfying the SCM with these possible additional assumptions on P_U, we have $\Psi^F(P_{U,X}) = \Psi(P(P_{U,X}))$ for some Ψ. O is implied by the distribution of (U, X), such as $O = X$ or $O \subset X$, and $P = P(P_{X,U})$, where $P(P_{U,X})$ is a distribution of O implied by $P_{U,X}$. Let us denote the resulting full-data SCM by $\mathcal{M}^{F*} \subset \mathcal{M}^F$ to make clear that possible additional assumptions were made that were driven purely by the identifiability problem, not necessarily reflecting reality. We now have that for each $P_{U,X} \in \mathcal{M}^{F*}$, $\Psi^F(P_{U,X}) = \Psi(P)$, with $P = P(P_{U,X})$ the distribution of O implied by $P_{U,X}$ (whereas P_0 is the true distribution of O implied by the true distribution $P_{U,X,0}$).

Theorems exist that are helpful to establish such a desired identifiability result. For example, if $O = X$, and the distribution of U is such that, for each s, A_s is independent of L_d, given $Pa(A_s)$, then the well-known g-formula expresses the distribution of L_d in terms of the distribution of O:

$$P(L_d = l) = \prod_{r=1}^{R} P(L_r = l_r \mid Pa_d(L_r)) = Pa_d(l_r)),$$

where $Pa_d(L_r)$ are the parents of L_r with the intervention nodes among these parent nodes deterministically set by intervention d.

This so-called sequential randomization assumption can be established for a particular independence structure of U by verifying the backdoor path criterion on the corresponding causal graph implied by the SCM and this independence structure on U. The backdoor path criterion states that for each A_s, each backdoor path from A_s to an L_r node that is realized after A_s is blocked by one of the other L_r nodes. One might be able to generate a number of independence structures on the distribution of U that provide the desired identifiability result. That is, the resulting model for U that provides the desired identifiability might be represented as a union of models for U that assume a specific independence structure. If there is only one intervention node, i.e., $S = 1$, so that $O = (W, A, Y)$, the sequential randomization assumption reduces to the randomization assumption (also known as the no unmeasured confounders assumption). The randomization assumption states that treatment node A is independent of counterfactual Y_a, conditional on W: $Y_a \perp A \mid Pa(A) = W$. We note that different such subsets of X may provide a desired identifiability result.

If we return to our example and the structural equation models found in (2.1), the union of several independence structures allows for the identifiability of our

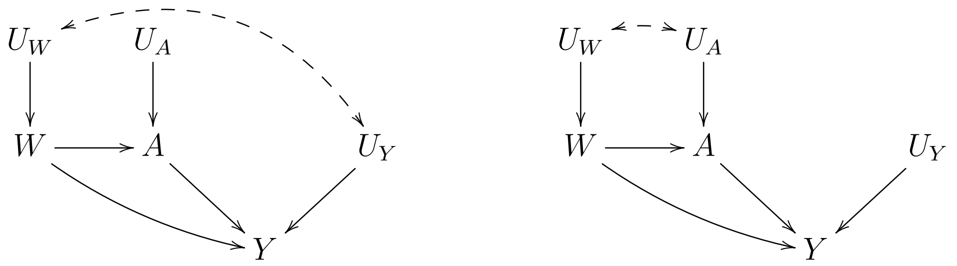


Figure 2.3: Causal graphs for (2.1) with various assumptions about the distribution of P_U

causal target parameter $E_0Y_1 - E_0Y_0$ by meeting the backdoor path criterion. The independence structure in Fig. 2.2 does not meet the backdoor path criterion, but the two in Fig. 2.3 do. Thus in these two graphs the randomization assumption holds: A and Y_a are conditionally independent given W, which is implied by U_A being independent of U_Y, given W. It should be noted that Fig. 2.1 is a special case of the first graph in Fig. 2.3, so the union model for the distribution of U only represents two conditional independence models.

Commit to a statistical model and target parameter. The identifiability result provides us with a purely statistical target parameter $\Psi(P_0)$ on the distribution P_0 of O. The full-data model $\mathcal{M}^{F*}$ implies a statistical observed data model $\mathcal{M} = \{P(P_{X,U}) : P_{X,U} \in \mathcal{M}^{F*}\}$ for the distribution $P_0 = P(P_{U,X,0})$ of O. This now defines a target parameter $\Psi : \mathcal{M} \to \mathbb{R}^d$. The statistical observed data model for the distribution of O might be the same for $\mathcal{M}^F$ and $\mathcal{M}^{F*}$. If not, then one might consider extending the Ψ to the larger statistical observed data model implied by $\mathcal{M}^F$, such as possibly a fully nonparametric model allowing for all probability distributions. In this way, if the more restricted SCM holds, our target parameter would still estimate the target parameter, but one now also allows the data to contradict the more restricted SCM based on additional doubtful assumptions.

We can return to our example and define our parameter, the causal risk difference, in terms of the corresponding statistical parameter $\Psi(P_0)$:

$$\Psi^F(P_{U,X,0}) = E_0Y_1 - E_0Y_0 = E_0[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W)] \equiv \Psi(P_0),$$

where the outer expectation in the definition of $\Psi(P_0)$ is the mean across the strata for W. This identifiability result for the additive causal effect as a parameter of the distribution P_0 of O required making the randomization assumption stating that A is independent of the counterfactuals (Y_0, Y_1) within strata of W. This assumption might have been included in the original SCM $\mathcal{M}^F$, but, if one knows there are unmeasured confounders, then the model $\mathcal{M}^{F*}$ would be more restrictive by enforcing this “known to be wrong” randomization assumption.

Another required statistical assumption is that $P_0(A = 1, W = w) > 0$ and

$P_0(A = 0, W = w) > 0$ are positive for each possible realization w of W. Without this assumption, the conditional expectations of Y in $\Psi(P_0)$ are not well defined.

To be very explicit about how this parameter corresponds with mapping P_0 into a number:

$$\begin{aligned}\Psi(P_0) \;&=\; \sum_w \Big[\sum_y y P_0(Y = y \mid A = 1, W = w) \\ &\quad - \sum_y y P_0(Y = y \mid A = 0, W = w) \Big] P_0(W = w),\end{aligned}$$

where

$$P_0(Y = y \mid A = a, W = w) \;=\; \frac{P_0(W = w, A = a, Y = y)}{\sum_y P_0(W = w, A = a, Y = y)}$$

is the conditional probability distribution of $Y = y$, given $A = a, W = w$, and

$$P_0(W = w) = \sum_{y,a} P_0(Y = y, A = a, W = w)$$

is the marginal probability distribution of $W = w$. This statistical parameter Ψ is defined on all probability distributions of (W, A, Y). The statistical model $\mathcal{M}$ is nonparametric and $\Psi : \mathcal{M} \to \mathbb{R}$.

Interpretation of target parameter. The observed data parameter $\Psi(P_0)$ can be interpreted in two possibly distinct ways:

1. $\Psi(P_0)$ with $P_0 \in \mathcal{M}$ augmented with the truly reliable additional nonstatistical assumptions that are known to hold (e.g., $\mathcal{M}^F$). This may involve bounding the deviation of $\Psi(P_0)$ from the desired target causal effect $\Psi^F(P_{U,X,0})$ under a realistic causal model $\mathcal{M}^F$ that is not sufficient for the identifiability of this causal effect.

2. The truly causal parameter $\Psi^F(P_{U,X}) = \Psi(P_0)$ under the more restricted SCM $\mathcal{M}^{F*}$, thereby now including all causal assumptions that are needed to make the desired causal effect identifiable from the probability distribution P_0 of O.

The purely statistical (noncausal) parameter given by interpretation 1 is often of interest, such as $E_{W,0}[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W)]$, which can be interpreted as the average of the difference in means across the strata for W. With this parameter we can assume nothing, beyond the positivity assumption, except perhaps time ordering $W \to A \to Y$, to have a meaningful interpretation of the difference in means. Since we do not assume an underlying system, the SCM for (U, X) and thereby Y_a, or the randomization assumption, the parameter is a statistical parameter only. This type of parameter is sometimes referred to as a variable importance measure.

2.2 Targeted Maximum Likelihood Learning

Targeted maximum likelihood super learning allows us to avoid reliance on unrealistic (parametric) statistical models, define interesting parameters, and target the fit of the data-generating distribution to the parameter of interest. The incorporation of super learning in the TMLE means that every effort is made to achieve minimal bias and the asymptotic semiparametric efficiency bound for variance. We discuss both procedures below.

Effect Estimation vs. Prediction. Both causal effect and prediction research questions are inherently *estimation* questions. In the first, we are interested in estimating the causal effect of A on Y adjusted for covariates W. For prediction, we are interested in generating a function to input the variables (A, W) and predict a value for Y. These are separate and distinct research questions. However, many (causal) effect estimators, such as TMLE, involve prediction steps within the procedure.

2.2.1 Super Learner

The first step in the TMLE is an initial estimate of the data-generating distribution P_0, or the relevant part that is needed to evaluate the target parameter. This is the first place where we will use super learner. An estimator maps the $O_1, \ldots, O_n$ observations into a value for the parameter it targets. We can view estimators as mappings from the empirical distribution P_n of the data set, where P_n places probability $1/n$ on each observed O_i, $i = 1, \ldots, n$. We want to use an algorithm to estimate the function $\bar{Q}_{X,0} : (A, W) \rightarrow \bar{Q}_{X,0}(A, W)$ (where $\bar{Q}_{X,0}(A, W) = E_{X,0}(Y \mid A, W)$). However, there are multiple "algorithms" to choose from, and we want to use the best estimator. We select this best estimator in terms of a loss function, which assigns a measure of performance to a candidate function $\bar{Q}$ when applied to an observation O. For binary Y, both the L_2 loss $L(O, \bar{Q}) = (Y - \bar{Q}(A, W))^2$ and negative log loss $L(O, \bar{Q}) = -\log(\bar{Q}(A, W)^Y (1 - \bar{Q}(A, W))^{1-Y})$ target the same function $\bar{Q}_{X,0}(A, W) = P_{X,0}(Y = 1 \mid A, W)$. We will use the L_2 loss because of its desirable boundedness properties.

We can now define $\bar{Q}_{X,0}(A, W) = E_{X,0}(Y \mid A, W)$ as the minimizer of the expected squared error loss:

$$\bar{Q}_{X,0} = \arg\min_{\bar{Q}} E_{X,0} L(O, \bar{Q}),$$

where $L(O, \bar{Q}) = (Y - \bar{Q}(A, W))^2$. $E_{X,0} L(O, \bar{Q})$, which we want to be small, evaluates the candidate $\bar{Q}$, and it is minimized at the optimal choice of $\bar{Q}_{X,0}$. We want the estimator of the regression function $\bar{Q}_{X,0}$ whose realized value minimizes the expectation of the squared error loss function.

How do we find out which algorithm among a collection or library of algorithms yields the smallest expected loss, or, equivalently, which one has the best performance with respect to the dissimilarity implied by the loss function? A library consists of various algorithms, such as random forests or parametric logistic regression. We can use these algorithms to build a library of algorithms consisting of all weighted averages of these regressions. It is reasonable to expect that one of these weighted averages might perform better than one of the regressions/algorithms alone. This simple principle allows us to map a collection of candidate algorithms into a library of weighted averages of these algorithms. Each weighted average is a unique candidate algorithm in this augmented library.

The entire data set (learning set) is fit using each of the algorithms in our collection of algorithms. The learning set is then split into V groups of size $\sim n/V$. We follow V-fold cross-validation, and for any given fold, $V-1$ sets will comprise the training set and the remaining 1 set is the validation set. The observations in the training set are used to construct (or train) the algorithms. The observations in the validation set are used to assess the performance (i.e., risk) of the candidate algorithms applied to the corresponding training set. The validation set rotates V times such that each set is used as the validation set once. Each algorithm is applied to the observations in the training set, and risk is estimated with the observations in the validation set. The risks obtained from the V validation sets are averaged to obtain the cross-validated risk for each algorithm. We then propose a family of weighted combinations of the algorithms and determine which combination minimizes the cross-validated risk over the family of weighted combinations. This family of weighted combinations includes only those that sum up to one and where each weight is positive or zero. The weighted combination with the smallest cross-validated risk is the best estimator according to our criteria: minimizing the estimated expected squared error loss function. (This cross-validated risk criterion can be applied to arbitrary loss functions.) The super learner returns a function that we can use for prediction in new data sets. We also use cross-validation to evaluate the overall performance of the super learner itself, by running the super learner within each of the V folds, and calculating a cross-validated risk.

Demonstrations of the super learner's superior finite sample performance in simulations and publicly available data sets, as well as asymptotic results, are discussed in van der Laan et al. (2007), Polley and van der Laan (2010), and van der Laan and Rose (2011). In brief, the asymptotic results prove that in realistic scenarios (where none of the algorithms are a correctly specified parametric statistical model), the cross-validated selector performs asymptotically as well as the oracle, which we define as the best estimator given the algorithms in the collection of algorithms. Consequently, super learner performs asymptotically as well as the best choice among the family of weighted combinations of estimators. Thus, by adding more competitors, we only improve the performance of the super learner. The asymptotic equivalence remains true if the number of algorithms in the library grows very fast with sample size.

2.2.2 TMLE

The TMLE is a general procedure for estimation of a target parameter of the data-generating distribution in semiparametric models. It marries the locally efficient double robust properties of estimating function based methodology and the properties of maximum likelihood estimation. TMLEs are loss-based well-defined, efficient, unbiased substitution estimators of the target parameter of the data-generating distribution. The estimator is a two-step procedure where one first obtains an estimate of the data-generating distribution P_0, or the relevant portion Q_0 of P_0. The second stage updates this initial fit in a step targeted toward making an optimal bias–variance tradeoff for the parameter of interest $\Psi(Q_0)$, instead of the density P_0.

Suppose that, given n i.i.d. observations $X_1, \ldots, X_n$, $P^*_{X,n}$ is a TMLE of $P_{X,0}$, and $\Psi^F(P^*_{X,n})$ is the corresponding TMLE of ψ^F_0. Specifically, let $L^F(P_X)(X)$ be a full-data loss function (e.g., log-likelihood loss function) so that

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

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

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

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

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

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

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

The TMLE will be explained in further detail in subsequent chapters in the context of CCW-TMLE. We also refer to Chapter 1 for additional literature on TMLE, and van der Laan and Rose (2011) for an expanded introduction to targeted learning. We will follow the road map for targeted learning (Fig. 2.4) in Chapters 3–5 with case-control data.

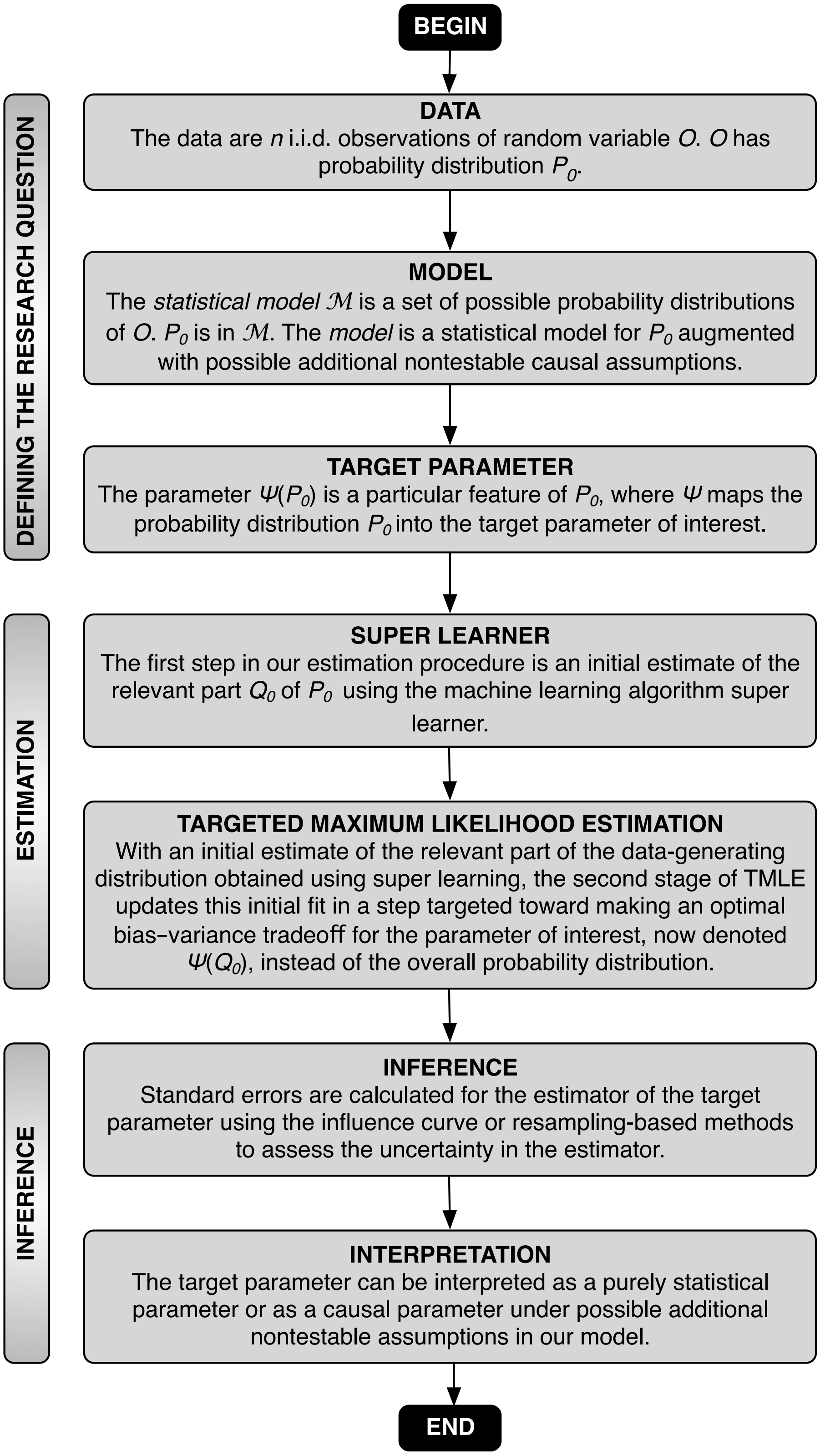

BEGIN
DEFINING THE RESEARCH QUESTION
DATA
The data are n i.i.d. observations of random variable O. O has probability distribution P_0.
MODEL
The *statistical model* $\mathcal{M}$ is a set of possible probability distributions of O. P_0 is in $\mathcal{M}$. The *model* is a statistical model for P_0 augmented with possible additional nontestable causal assumptions.
TARGET PARAMETER
The parameter $\Psi(P_0)$ is a particular feature of P_0, where Ψ maps the probability distribution P_0 into the target parameter of interest.
ESTIMATION
SUPER LEARNER
The first step in our estimation procedure is an initial estimate of the relevant part Q_0 of P_0 using the machine learning algorithm super learner.
TARGETED MAXIMUM LIKELIHOOD ESTIMATION
With an initial estimate of the relevant part of the data-generating distribution obtained using super learning, the second stage of TMLE updates this initial fit in a step targeted toward making an optimal bias–variance tradeoff for the parameter of interest, now denoted $\Psi(Q_0)$, instead of the overall probability distribution.
INFERENCE
INFERENCE
Standard errors are calculated for the estimator of the target parameter using the influence curve or resampling-based methods to assess the uncertainty in the estimator.
INTERPRETATION
The target parameter can be interpreted as a purely statistical parameter or as a causal parameter under possible additional nontestable assumptions in our model.
END

Figure 2.4: Road map for targeted learning

Chapter 3

Targeted Learning in Independent Case-Control Designs

Our proposed CCW-TMLE for case-control studies targets the parameter of interest and 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 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.

3.1 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 control status. Our target parameter of interest might be the causal risk difference, which we 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

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

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

where the cluster containing one case and J controls is considered the experimental unit. 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 the formulation as described here does not describe *individually matched* case-control sampling, which we describe in Chapter 4.

3.2 CCW-TMLE

We discuss a CCW-TMLE for the causal risk difference with $X = (W, A, Y) \sim P_{X,0}$ and $O = \big((W_1, A_1), (W_0^j, A_0^j : j = 1, \ldots, J)\big) \sim P_0$. The full-data efficient influence curve $D^F(Q_0, g_0)$ at $P_{X,0}$ is given by

$$\begin{aligned} D^F(Q_0, g_0) &= \left(\frac{I(A = 1)}{g_0(1 \mid W)} - \frac{I(A = 0)}{g_0(0 \mid W)}\right)(Y - \bar{Q}_0(A, W)) \\ &\quad + \bar{Q}_0(1, W) - \bar{Q}_0(0, W) - \Psi^F(Q_0), \end{aligned} \tag{3.1}$$

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

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

3.2.1 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.

3.2.2 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.

3.2.3 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 $Q^*_n = Q^1_n$.

3.2.4 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:

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

3.2.5 Calculating Standard Errors

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 (3.1). The case-control-weighted influence curve for the risk difference is then estimated by

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

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 $\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} IC_n^2(O_i)$.

3.3 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 this chapter and in van der Laan (2008a) and Rose and van der Laan (2008).

IPTW estimator. Presented in Chapter 1, Robins (1999) 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. Mean squared errors (MSEs) and relative efficiencies (REs) for the causal odds ratio are provided in Table 3.1. 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. 4.1. The CCW-MLE and CCW-TMLE with misspecified initial $\bar{Q}_n(A, W)$ had more bias than their correctly specified counterparts.

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 $q_0 = 0.053$, and 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 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

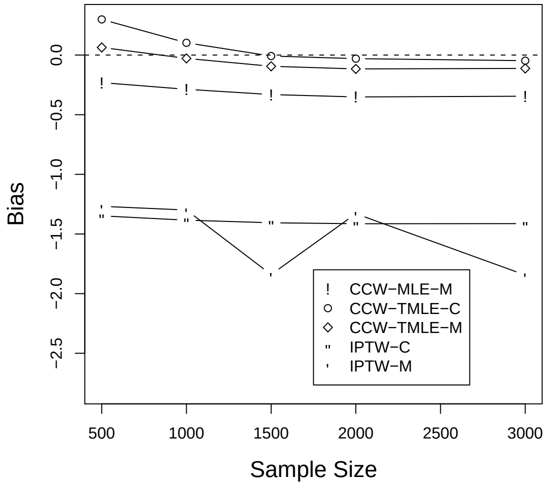


Figure 3.1: Simulation 1 bias results. *Bias results for the CCW-TMLE with misspecified $g_n(A \mid W)$ and the correctly specified CCW-MLE were excluded since values were the same as those for the TMLE with correctly specified $\bar{Q}_n(A, W)$ and $g_n(A \mid W)$.*

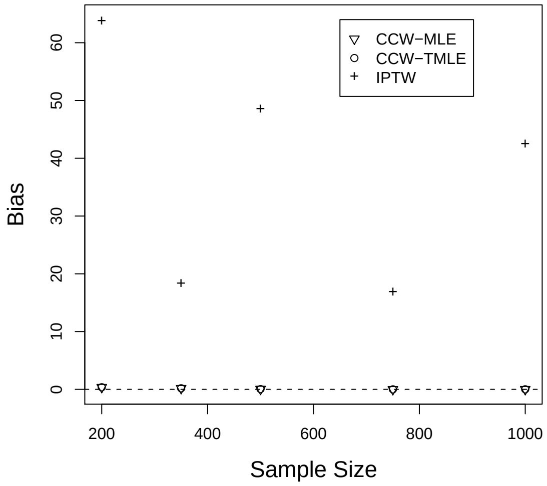


Figure 3.2: Simulation 2 bias results. *Bias results for the CCW-TMLE with misspecified $g_n(A \mid W)$ were excluded since those values were the same as those for the CCW-TMLE with correctly specified $\bar{Q}_n(A, W)$ and $g_n(A \mid W)$*

Table 3.1: Results for the odds ratio in simulations 1 and 2. *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 $g_n(A \mid W)$*

		nC	250	500	500	1000	1000
Simulation 1	Fit	nCo	250	500	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

		nC	100	250	250	500
Simulation 2	Fit	nCo	250	250	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

Table 3.2: Standard error illustration in simulation 2. *OR is odds ratio, SE is standard error, CI is confidence interval, p is p-value. 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

the results here are more extreme. Results can be seen in Table 3.1 and Fig. 3.2.

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 CCW-TMLE. We sampled one data set with size $n = 1000$ 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 $OR = 3.42$. Standard error estimates for the IPTW estimator were calculated by bootstrapping the case and control samples 1000 times. The results are presented in Table 3.2, including odds ratio estimates, standard errors, confidence intervals, and p-values. Here, we compare only the CCW-TMLE and the IPTW estimator. (CCW-MLE was excluded as we wished 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.) The IPTW estimators are more biased and considerably less efficient than the CCW-TMLEs.

Remark: intercept-adjusted maximum likelihood estimation. intercept-adjusted maximum likelihood estimation, discussed in Chapter 1, and case-control-weighted maximum likelihood estimation are both options for the initial fit $\bar{Q}_n(A, W)$. Issues became apparent when using intercept-adjusted maximum likelihood estimation in our CCW-TMLE framework. In multiple simulation settings, we found that when $\bar{Q}_n(A, W)$ was misspecified using an intercept-adjusted fit, the predicted probabilities were substantially biased compared to the misspecified case-control-weighted maximum likelihood probabilities. This additional bias can be understood intuitively since the update to the logistic regression is static regardless of the specification used, and the parameters of the regression (excluding the intercept) are not adjusted by this update. For a correctly specified initial fit this is not a problem, but when $\bar{Q}_n(A, W)$ is realistically 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 $\bar{Q}_n(A, W)$, case-control-weighted predicted probabilities will likely be closer to the truth than intercept-adjusted estimates. See Figure 3.3 for an illustration.

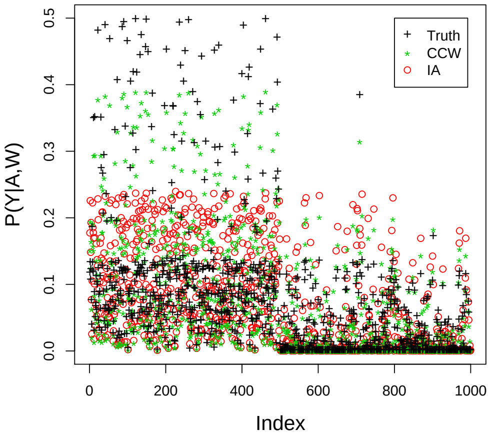


Figure 3.3: Example: Predicted probabilities for a misspecified $\bar{Q}_n(A, W)$

CCW-TMLE using an intercept-adjusted initial fit improved, with regard to bias, on its non-targeted counterpart for misspecified $\bar{Q}_n(A, W)$. However, the additional bias for misspecified $\bar{Q}_n(A, W)$ and intercept-adjusted logistic regression led to much slower convergence to the true values of the risk difference, relative risk, and odds ratio. CCW-TMLE with misspecified $\bar{Q}_n(A, W)$ fit with case-control-weighted logistic regression became consistent for reasonable sample sizes. Coverage probabilities for CCW-TMLE using an intercept-adjusted initial fit for misspecified $\bar{Q}_n(A, W)$ also diverged substantially from 95% (as low as 65%) for reasonable sample sizes due to the bias of the estimators. When $\bar{Q}_n(A, W)$ was correctly specified, the intercept-adjusted methods performed as well as the case-control-weighted methods. However, since correct specification of $\bar{Q}_n(A, W)$ is unlikely in practice, this is a significant drawback, and we did not include intercept adjustment in our simulations presented above.

Simulation 3. Our third simulation study was designed to illustrate the performance of CCW-TMLE when q_0 is estimated. We examine MSE, coverage probabilities, and percentage of rejected tests. 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: $W \sim U(0, 1)$, $P_{X,0}(A = 1 \mid W) = \text{expit}(W^2 - 4W + 1)$, and $P_{X,0}(Y = 1 \mid A, W) = \text{expit}(A - \sin W^2 + A \sin W^2 + 7A \times \log W + 5 \log W - 1)$. 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 estimated prevalence probability $q_n = (0.02, 0.03, 0.04)$. The true value for the odds ratio was given by $OR = 1.851$. As in previous simulations, we used both correctly specified and misspecified fits for $\bar{Q}_n(A, W)$ and $g_n(A \mid W)$. The misspecified $\bar{Q}_n(A, W)$ included A and W as main terms and the misspecified $g_n(A \mid W)$ included only W.

When examining the MSE results for the odds ratio across the range of values for q_n, 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 preliminary evidence that the CCW-TMLE 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. Results are displayed in Table 3.3. 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 of van der Laan (2008a). This theorem demonstrates that one can incorporate the standard error of the estimate q_n into the confidence interval for the parameter of interest.

Table 3.3: Results for the odds ratio in simulation 3. *CP is coverage probability, and %RT is percent rejected tests (*$\alpha = 0.05$*). Results are for 1000 simulations of 1000 individuals with 500 cases and 500 controls randomly sampled from the population in Simulation 3, true* $OR = 1.851$

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

3.4 Discussion

Case-control weighting provides a framework for the analysis of case-control study designs using TMLEs. We observed that the "approximately correct" IPTW method was outperformed by CCW-TMLE under conditions similar to a practical setting 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-TMLEs. Further, in practice we recommend the use of super learner for the estimation of $\bar{Q}_0$. We also introduced the CCW-MLE, which provides an alternative for practitioners without the statistical support to implement the CCW-TMLE. 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.

Chapter 4

Targeted Learning in Individually Matched Case-Control Designs

Individually matched case-control study designs are common in public health and medicine, and conditional logistic regression in a parametric statistical model is the tool most commonly used to analyze these studies. In an individually matched case-control study, the population of interest is identified, and cases are randomly sampled. Each of these cases is then matched to one or more controls based on a variable (or variables) *believed* to be a confounder. The main potential benefit of matching in case-control studies is a gain in efficiency, not the elimination of confounding. Therefore, when are these study designs truly beneficial? Given the potential drawbacks, including extra cost, added time for enrollment, increased bias, and potential loss in efficiency, the use of matching in case-control study designs warrants careful evaluation.

In this chapter, we focus on individual matching in case-control studies where the researcher is interested in estimating a causal effect and certain prevalence probabilities are known or estimated. In order to eliminate the bias caused by the matched case-control sampling design, this technique relies on knowledge of the true prevalence probability $q_0 \equiv P_{X,0}(Y = 1)$ and an additional value:

$$\bar{q}_0(M) \equiv q_0 \frac{P_{X,0}(Y = 0 \mid M)}{P_{X,0}(Y = 1 \mid M)},$$

where M is the matching variable. We will compare the use of CCW-TMLEs in matched and unmatched case-control study designs as we explore which design yields the most information for the causal effect of interest.

4.1 Data, Model, and Target Parameter

We define $X = (W, M, A, Y) \sim P_{X,0}$ as the experimental unit and corresponding distribution $P_{X,0}$ of interest. Here X consists of baseline covariates W, an exposure variable A, and a binary outcome Y, which defines case or control status. We can define $\psi_0^F = \Psi^F(P_{X,0}) \in \mathbb{R}^d$ of $P_{X,0} \in \mathcal{M}^F$ as the causal effect parameter, and for binary exposure $A \in \{0,1\}$ we define the risk difference, relative risk, and odds ratio as in the previous chapter. The observed data structure in matched case-control sampling is defined by

$$O = \big((M_1, W_1, A_1), (M_0^j = M_1, W_0^j, A_0^j : j = 1, \ldots, J)\big) \sim P_0, \text{ with}$$

$$(M_1, W_1, A_1) \sim (M, W, A \mid Y = 1) \text{ for cases and}$$

$$(M_0^j, W_0^j, A_0^j) \sim (M, W, A \mid Y = 0, M = M_1) \text{ for controls.}$$

Here $M \subset W$, and M is a categorical matching variable. The sampling distribution of data structure O is described as above with P_0. Thus, the matched case-control data set contains n independent and identically distributed observations $O_1, \ldots, O_n$ with sampling distribution P_0. The cluster containing one case and the J controls is the experimental unit, and the marginal distribution of the cluster is specified by the population distribution $P_{X,0}$. The model $\mathcal{M}^F$, which possibly includes knowledge of q_0 or $\bar{q}_0(M)$, then implies models for the probability distribution of O consisting of cases (M_1, W_1, A_1) and controls $(M_1, W_2^j, A_2^j), j = 1, \ldots, J$.

4.2 CCW-TMLE for Individual Matching

CCW-TMLEs for individually matched case-control studies incorporate knowledge of q_0 and $\bar{q}_0(M)$, where $\bar{q}_0(M)$ is defined as

$$\bar{q}_0(M) \equiv q_0 \frac{P_{X,0}(Y = 0 \mid M)}{P_{X,0}(Y = 1 \mid M)} = q_0 \frac{q_0(0 \mid M)}{q_0(1 \mid M)}.$$

Implementation of CCW-TMLE in individually matched studies echos the procedure for independent (unmatched) case-control studies, with the exception that the weights now differ. We summarize this procedure assuming the knowledge presented in the previous chapter. We focus on the risk difference $\psi_{RD,0}^F = E_{X,0}[E_{X,0}(Y \mid A = 1, W) - E_{X,0}(Y \mid A = 0, W)]$ as an illustrative example.

Implementing CCW-TMLE for Individually Matched Data

Step 0. Assign weights q_0 to cases and $\bar{q}_0(M)/J$ to the corresponding J controls.

Step 1. Estimate the conditional probability of Y given A and W using super learning and assigned weights. The estimate of $P_{X,0}(Y = 1 \mid A, W, M) \equiv \bar{Q}_0(A, W, M)$ is $\bar{Q}_n^0(A, W, M)$. Let Q_n^0 be the estimate of the conditional mean and the case-control-weighted empirical distribution for the marginal distribution of W, representing the estimator of $Q_0 = (\bar{Q}_0, Q_{W,0})$.

Step 2. Estimate the exposure mechanism using super learning and weights. The estimate of $P_{X,0}(A \mid W, M) \equiv g_0(A \mid W, M)$ is $g_n(A \mid W, M)$.

Step 3. Determine a parametric family of fluctuations $Q_n^0(\epsilon)$ of Q_n^0 with fluctuation parameter ϵ, and a case-control-weighted loss function $L_{q_0}(Q) = q_0 L^F(Q)(M_1, W_1, A_1, 1) + (\bar{q}_0(M)/J) \sum_{j=1}^J L^F(Q)(M_1, W_2^j, A_2^j, 0)$ such that the derivative of $L^F(Q_n^0(\epsilon))$ at $\epsilon = 0$ equals the full-data efficient influence curve at any initial estimator $Q_n^0 = (\bar{Q}_n^0, Q_{W,n}^0)$ and g_n. Since initial Q_{Wn}^0 is the empirical distribution (i.e., case-control-weighted nonparametric maximum likelihood estimation), one only needs to fluctuate $\bar{Q}_n^0$ and the fluctuation function involves a choice of clever covariate chosen such that the above derivative condition holds. Calculate the clever covariate $H_n^*(A, W, M)$ for each subject as a function of $g_n(A \mid W, M)$:

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

Step 4. Update the initial fit $\bar{Q}_n^0(A, W, M)$ from step 1 using the covariate $H_n^*(A, W, M)$. This is achieved by holding $\bar{Q}_n^0(A, W, M)$ fixed while estimating the coefficient ϵ for $H_n^*(A, W, M)$ in the fluctuation function using case-control-weighted maximum likelihood estimation. Let ϵ_n be this case-control-weighted parametric maximum likelihood estimator. The updated regression is given by $\bar{Q}_n^1 = \bar{Q}_n^0(\epsilon_n)$. No iteration is necessary since the next ϵ_n will be equal to zero. The CCW-TMLE of Q_0 is now $Q_n^* = (\bar{Q}_n^1, Q_{Wn}^0)$, where only the conditional mean estimator $\bar{Q}_n^0$ was updated.

Step 5. Obtain the substitution estimator of the target parameter by application of the target parameter mapping to Q_n^*:

$$\psi_n^* = \left\{ \frac{1}{n} \sum_{i=1}^n \left(q_0 \bar{Q}_n^1(1, W_{1,i}, M_{1,i}) + \frac{\bar{q}_0(M)}{J} \sum_{j=1}^J \bar{Q}_n^1(1, W_{2,i}^j, M_{1,i}) \right) - \left(q_0 \bar{Q}_n^1(0, W_{1,i}, M_{1,i}) + \frac{\bar{q}_0(M)}{J} \sum_{j=1}^J \bar{Q}_n^1(0, W_{2,i}^j, M_{1,i}) \right) \right\}.$$

Step 6. Calculate standard errors, p-values, and confidence intervals based on the influence curve of the CCW-TMLE ψ_n^*. The influence curve can be selected to be the case-control-weighted full-data efficient influence curve (just as we defined the case-control-weighted full-data loss function).

4.3 Simulations

In the following simulation studies, we compare the CCW-TMLE in independent and individually matched study designs.

Simulation 1. Our first simulation study is designed to illustrate the differences between independent case-control sampling and matched case-control sampling in "ideal" situations where control information is not discarded (e.g., data collection is expensive, and covariate information is only collected when a control is a match). This simulation also demonstrates the use of weights q_0 and $(1-q_0)\frac{1}{J}$ with matched data, to represent situations where $\bar{q}_0(M)$ is not known. The population contained $N = 35{,}000$ individuals, where we simulated a 9-dimensional covariate $W = (W_i : i = 1, \ldots, 9)$, a binary exposure (or "treatment") A, and an indicator Y. These variables were generated according to the following rules: $P_{X,0}(W_i = 1) = 0.5$, $P_{X,0}(A = 1 \mid W) = \text{expit}(W_1 + W_2 + W_3 - 2W_4 - 2W_5 + 2W_6 - 4W_7 - 4W_8 + 4W_9)$, and $P_{X,0}(Y = 1 \mid A, W) = \text{expit}(1.5A + W_1 - 2W_2 - 4W_3 - W_4 - 2W_5 - 4W_6 + W_7 - 2W_8 - 4W_9)$.

Both the exposure mechanism and the conditional mean of Y given its parents were generated with varied levels of association with A and Y in order to investigate the role of weak, medium, and strong association between a matching variable W_i and A and Y. The corresponding associations can be seen in Table 4.1. For example, W_1 was weakly associated with both A and Y. Matching is only potentially beneficial when the matching variable is a true confounder.

Table 4.1: Simulated covariates

			Y	
	Association	Weak	Medium	Strong
	Weak	W_1	W_2	W_3
A	Medium	W_4	W_5	W_6
	Strong	W_7	W_8	W_9

Another illustration of the varied association levels can be seen in Table 4.2, where we display the probability an individual in the population was a case given $W_i = w$, all the nonmatching covariates (Z), and A. For example, let's say matching variable W_2 is *age* with 1 representing <50 years old and 0 representing ≥ 50 years old. In this population, it was not very likely (0.013) that someone who is < 50 years old will become a case, while someone who is ≥ 50 years old has a much higher chance of becoming a case (0.047), given Z and A. Therefore, W_2, W_5, and W_8 represent situations where the distribution of W_i among cases and controls is very different. The covariates W_3, W_6, and W_9 represent situations where this difference is even more extreme.

The simulated population had a prevalence probability of $q_0 = 0.030$ and exactly 1,045 cases, and the true value of the odds ratio was given by $OR = 2.302$. We sampled the population using a varying number of cases $nC = (200,\ 500, 1000)$

in both matched and unmatched designs, and for each sample size we ran 1000 simulations. In each sample, the same cases were used for both designs. Controls were matched to cases in our matched simulations based on one variable (W_i) for both 1:1 and 1:2 designs. The causal odds ratio was estimated using a CCW-TMLE with correctly specified case-control-weighted logistic regressions.

The matched and unmatched designs performed similarly with respect to bias for the nine covariates, as shown in Figs. 4.1 and 4.2. There were consistent increases in efficiency when the association between W_i and Y was high (W_3, W_6, and W_9), when comparing matched to independent. Results when the association with W_i and Y was medium (W_2, W_5, and W_8) were not entirely consistent, although covariates W_5 and W_8 did show increases in efficiency for the matched design for all or nearly all sample sizes. These results are in line with the consensus found in the literature: that matching may produce gains in efficiency when the distribution of the matching variable differs drastically between the cases and the controls. Efficiency results for the odds ratio can be seen in Table 4.3.

Simulation 1 also demonstrates the use of weights q_0 and $(1-q_0)\frac{1}{J}$ with matched data, for situations where $\bar{q}_0(M)$ is unknown. This weighting scheme provided a reasonable approximation, yielding larger standard errors, but similar levels of bias for covariates with a weak association with Y. As association with Y increased, the estimate of the odds ratio became more biased. Bias results are presented in Figs. 4.1 and 4.2 and efficiency results are presented in Rose and van der Laan (2009).

Simulation 2. The second simulation study was designed to address less ideal more common situations where control information is discarded. Controls were sampled from the population of controls in simulation 1 until a match on covariate W_i was found for each case. Nonmatches were returned to the population of controls. The number of total controls sampled to find sufficient matches was recorded for each simulation. This was the number of randomly sampled controls that was used for the corresponding independent case-control simulation. The mean number of controls sampled to achieve 1:1 and 1:2 matching at each sample size is noted in Table 4.4 as nCo. For example, in order to obtain 200 controls matched on covariate W_1 in a 1:1 design, an average of 404 controls had to be sampled from the population. Thus, an average of 404 controls were used in the corresponding independent design.

CCW-TMLE was performed for both designs with correctly specified case-control-weighted logistic regression estimators for the exposure mechanism and conditional mean of Y given A and W. The independent design outperformed the matched design with respect to efficiency and bias for all sample sizes and both 1:1 and 1:2 matching. This was not surprising given the mean number of controls in each of the independent unmatched designs was, on average, about two times the number of controls for the matched design. Additionally, as association between W_i and Y increased, there was a trend that the number of controls necessary for complete matching also increased. A similar trend between A and W_i was not apparent. Bias results do not vary greatly with association between W_i and A or Y. Efficiency results can be seen in Table 4.4. Bias results are displayed in Fig. 4.3.

Table 4.2: Simulated covariates: probabilities

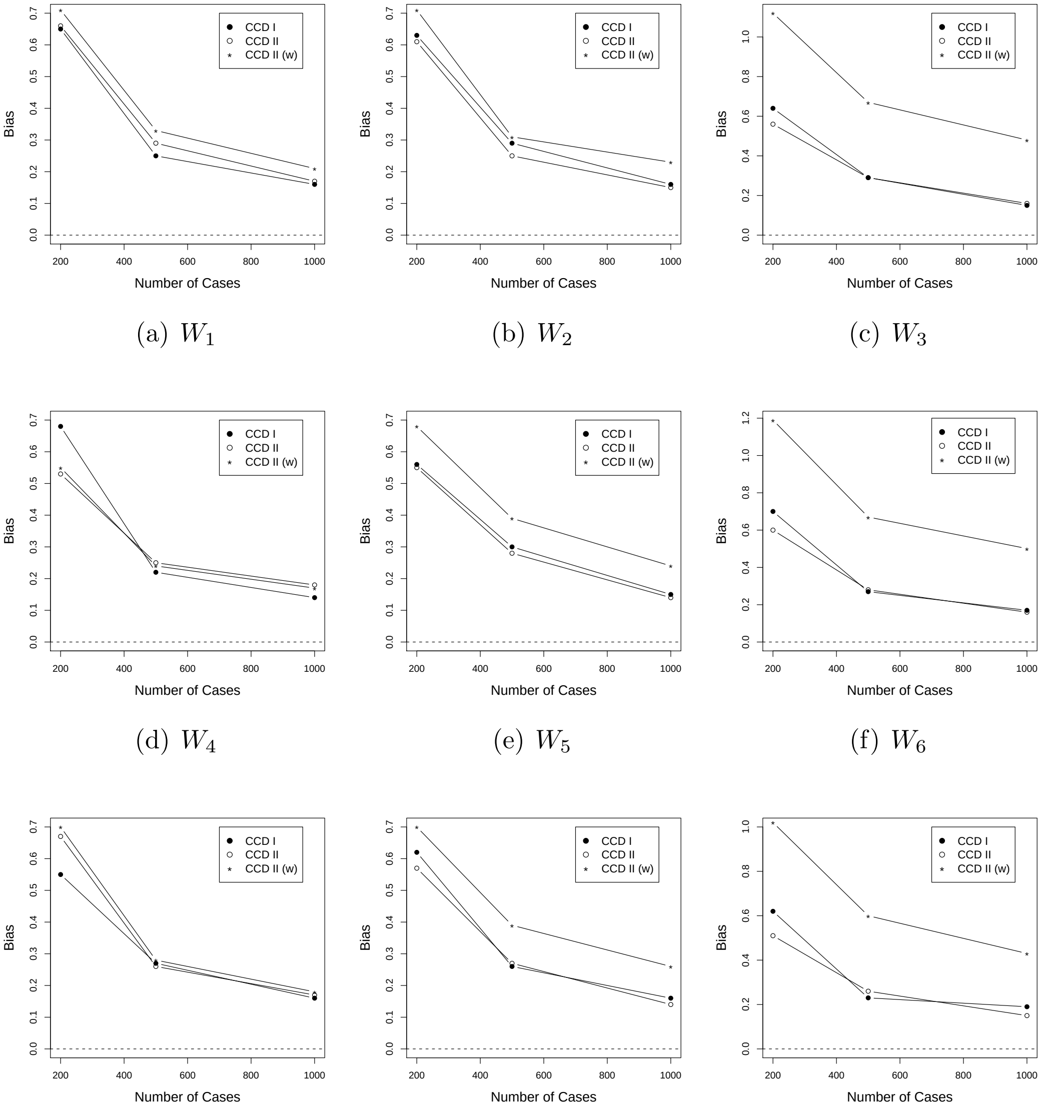
W_i	$P_{X,0}(Y = 1 \mid W_i = 1, Z, A)$	$P_{X,0}(Y = 1 \mid W_i = 0, Z, A)$
W_1	0.039	0.021
W_2	0.013	0.049
W_3	$\boxed{0.003}$	0.060
W_4	0.021	0.040
W_5	0.013	0.047
W_6	$\boxed{0.003}$	0.061
W_7	0.040	0.023
W_8	0.013	0.046
W_9	$\boxed{0.004}$	0.066

Table 4.3: Results for the odds ratio in simulation 1. *nC is number of cases*

			1:1			**1:2**	
	nC	200	500	1000	200	500	1000
W_1	Matched MSE	2.67	0.77	0.30	0.98	0.32	0.14
	Independent RE	1.09	1.05	1.03	0.97	0.97	1.00
W_2	Matched MSE	2.63	0.70	0.33	1.07	0.40	0.15
	Independent RE	1.01	0.93	1.18	1.00	1.21	1.07
W_3	Matched MSE	1.95	0.59	0.23	0.93	0.29	0.13
	Independent RE	0.80	0.78	0.79	0.90	0.88	1.00
W_4	Matched MSE	2.20	0.64	0.30	1.05	0.32	0.14
	Independent RE	0.77	1.07	1.11	1.00	0.94	0.93
W_5	Matched MSE	2.10	0.61	0.28	0.98	0.30	0.14
	Independent RE	0.82	0.80	0.93	0.91	0.83	1.00
W_6	Matched MSE	2.28	0.61	0.24	0.92	0.27	0.12
	Independent RE	0.74	0.97	0.80	0.95	0.84	0.86
W_7	Matched MSE	2.55	0.69	0.30	1.08	0.32	0.16
	Independent RE	1.11	0.96	1.00	0.98	1.00	1.23
W_8	Matched MSE	2.00	0.61	0.22	0.86	0.25	0.11
	Independent RE	0.78	0.88	0.76	0.90	0.78	0.85
W_9	Matched MSE	1.77	0.58	0.24	0.71	0.24	0.12
	Independent RE	0.72	0.91	0.77	0.63	0.75	0.92

Table 4.4: Results for the odds ratio in simulation 2. *nC is number of cases and nCo is mean number of controls for the independent case-control design*

			1:1			**1:2**	
	nC	200	500	1000	200	500	1000
	nCo	404	1006	2010	804	2011	4026
W_1	Matched MSE	2.90	0.76	0.28	1.00	0.27	0.14
	Independent RE	2.89	2.24	2.14	2.12	1.70	2.16
	nCo	404	1009	2016	808	2016	4031
W_2	Matched MSE	2.91	0.77	0.30	1.15	0.36	0.16
	Independent RE	2.91	2.72	2.13	2.32	2.21	2.49
	nCo	406	1016	2033	812	2034	4065
W_3	Matched MSE	1.99	0.48	0.22	0.84	0.28	0.11
	Independent RE	1.82	1.43	1.65	1.81	1.78	1.85
	nCo	403	1006	2010	806	2012	4023
W_4	Matched MSE	2.47	0.67	0.29	1.09	0.28	0.13
	Independent RE	2.38	2.09	2.20	2.29	1.91	2.03
	nCo	406	1010	2019	810	2019	4040
W_5	Matched MSE	2.41	0.63	0.25	0.92	0.29	0.12
	Independent RE	2.24	2.00	1.92	1.95	1.89	2.10
	nCo	411	1025	2046	819	2045	4094
W_6	Matched MSE	2.08	0.64	0.23	0.88	0.27	0.13
	Independent RE	2.13	1.99	1.69	1.92	1.70	2.23
	nCo	402	1001	2000	801	1999	4000
W_7	Matched MSE	2.71	0.72	0.30	1.09	0.34	0.15
	Independent RE	2.54	2.42	2.18	2.19	2.25	2.18
	nCo	407	1014	2028	811	2027	4055
W_8	Matched MSE	2.28	0.56	0.23	0.97	0.25	0.11
	Independent RE	2.35	1.76	1.71	1.99	1.59	1.68
	nCo	413	1030	2059	824	2061	4121
W_9	Matched MSE	1.97	0.54	0.22	0.80	0.26	0.12
	Independent RE	1.91	1.77	1.69	1.62	1.69	1.84

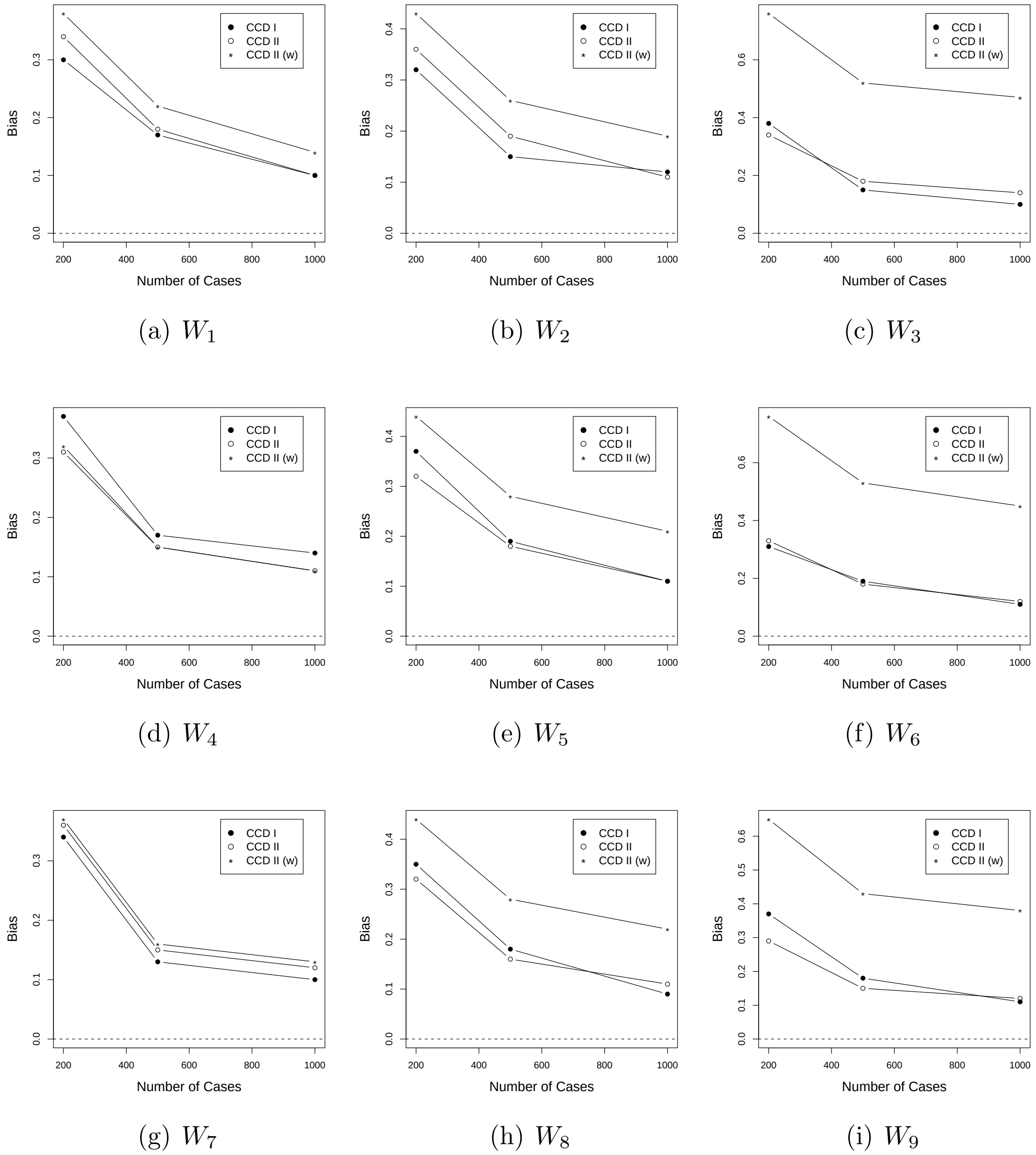


(a) W_1 (b) W_2 (c) W_3

(d) W_4 (e) W_5 (f) W_6

(g) W_7 (h) W_8 (i) W_9

Figure 4.1: Simulation 1 bias for 1:1 matching. *CCD I is "Case-Control Design I" referring to the independent case-control design, CCD II is "Case-Control Design II" referring to the matched case-control design with $\bar{q}_0(M)$ weighting, and CCD II (w) is the matched design with $(1 - q_0)$ weighting*

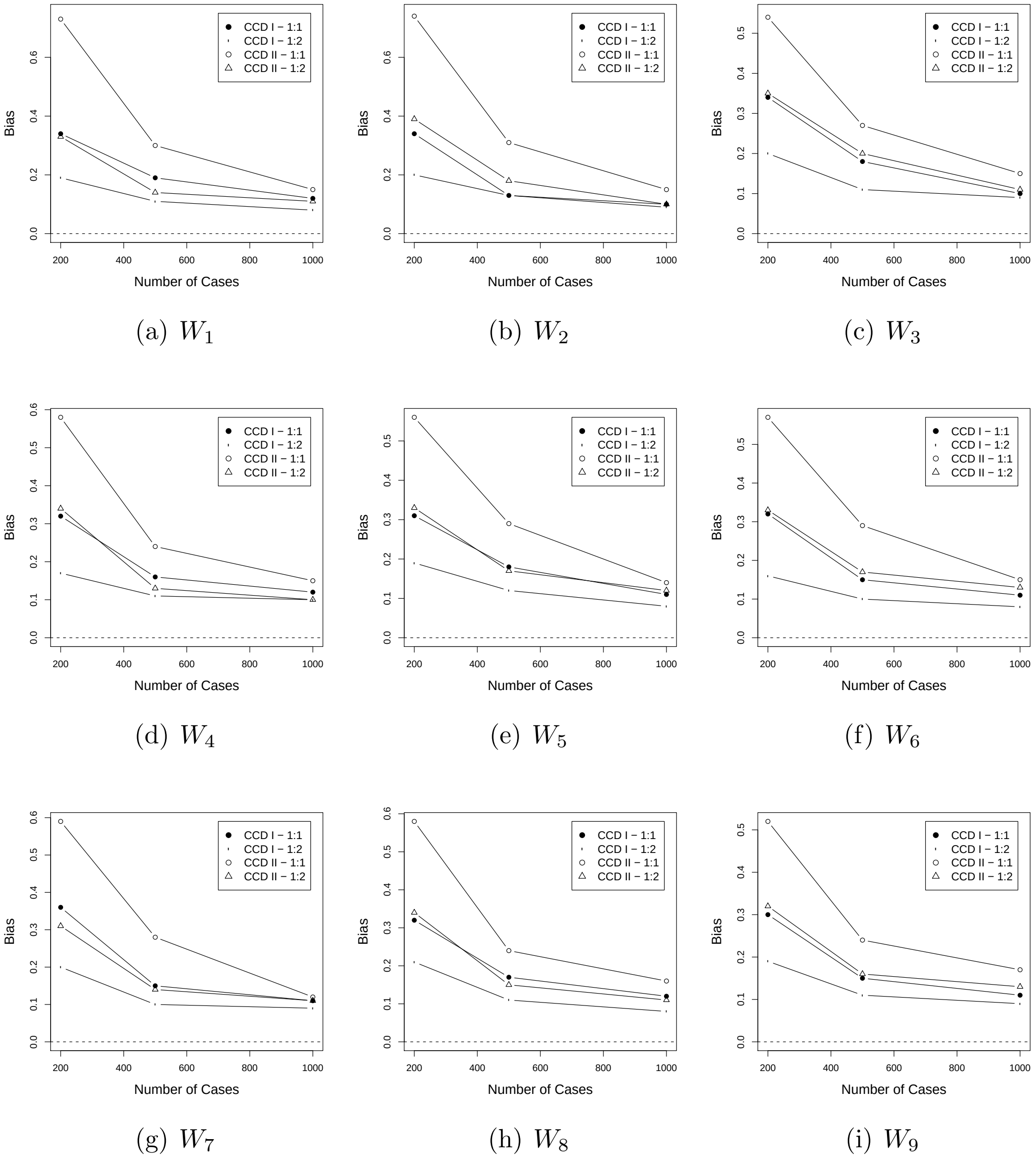


(a) W_1 (b) W_2 (c) W_3

(d) W_4 (e) W_5 (f) W_6

(g) W_7 (h) W_8 (i) W_9

Figure 4.2: Simulation 1 bias for 1:2 matching. *CCD I is "Case-Control Design I" referring to the independent case-control design, CCD II is "Case-Control Design II" referring to the matched case-control design with $\bar{q}_0(M)$ weighting, and CCD II (w) is the matched design with $(1 - q_0)$ weighting*



(a) W_1 (b) W_2 (c) W_3

(d) W_4 (e) W_5 (f) W_6

(g) W_7 (h) W_8 (i) W_9

Figure 4.3: Simulation 2 bias. *CCD I is "Case-Control Design I" referring to the independent case-control design and CCD II is "Case-Control Design II" referring to the matched case-control design*

4.4 Discussion

The main benefit of a matched case-control study design is a potential increase in efficiency. However, an increase in efficiency is not automatic. If one decides to implement a matched case-control study design, selection of the matching variable is crucial. Numerous publications in the literature indicate that matching on non-confounding variables is not beneficial, including Kupper et al. (1981): *"The futility of matching in [nonconfounding situations] is clear...matching on [the variable] will have absolutely no effect on the distribution of the exposure variable in the diseased and nondiseased groups."* In practice, it may be difficult to ascertain the strength of the association between the matching variable, the exposure of interest, and the outcome.

Our simulations confirmed the consensus in the existing literature: that in situations where the distribution of the matching covariate is drastically different between the case and control populations, matching may provide an increase in efficiency. Our simulations indicated that $P_{X,0}(Y = 1 \mid W_i = 1, Z, A)$, for matching variable W_i and covariate vector Z, may need to be very small for an increase in efficiency using a matched design. These results were true, however, only for simulations where *no control subjects were discarded*; it is very common for matched study designs to discard controls (Freedman 1950; Cochran 1965; Billewicz 1965; McKinlay 1977). We showed that in practical situations (e.g., when controls are discarded), an unmatched design is likely to be a more efficient, less biased study design choice. Our simulations also indicated that when $\bar{q}_0(M)$ is unknown, $(1 - q_0)\frac{1}{J}$ may provide a reasonable approximation, although this should be examined further. We did not address matching in cohort studies, and concentrated solely on case-control studies. However, matching in cohort studies was briefly addressed in van der Laan (2008a), and applying our methods in cohort studies is an area of future research.

Chapter 5

Targeted Learning in Two-Stage Designs

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

5.1 Effect Estimation

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

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

5.1.1 IPCW-TMLE in Two-Stage Samples

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

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

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

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

5.1.2 IPCW-TMLE

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

IPCW initial estimator. Computing an initial IPCW-loss based estimator $P_{X,n}^0$ (e.g., using super learning) based on, for example, the IPCW-loss function

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

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

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

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

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

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

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

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

IPCW Full-Data Efficient Influence Curve Equation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Risk Difference Example

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

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

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

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

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

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

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

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

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

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

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

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

Parametric submodel for full-data TMLE. Let

$$Q_{W,n}^0(\epsilon_1) = (1 + \epsilon_1 D_W^F(Q_n^0))Q_{W,n}^0$$

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

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

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

The IPCW-TMLE. Define

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

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

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

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

This is the IPCW-TMLE of ψ_0^F.

Right Censoring

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

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

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

Effect Modification

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

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

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

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

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

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

5.1.3 Simulations

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

SPPARCS Simulations

The National Institute of Aging funded Study of Physical Performance and Age-Related Changes in Sonomans (SPPARCS) is a population-based, census-sampled, study of the epidemiology of aging and health (Tager et al. 1998). 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. LTPA was calculated from answers to a detailed questionnaire where performed vigorous physical activities are assigned standardized intensity values in metabolic equivalents (METs). The recommended level of energy expenditure for the elderly is 22.5 METs.

The full-data structure is $X = (W, A, Y)$, where $Y = I(T \leq 5 \text{ years})$, T is time to the event death, A is a binary categorization of LTPA, and W are potential confounders. These variables are further defined in Table 5.1. The observed data structure on a randomly sampled subject can be represented as $O = (V, \Delta, \Delta X)$, where V is in X. Of note is the lack of any right censoring in this longitudinal cohort. The outcome (death within or at five years after baseline interview) and date of death was recorded for each subject. This information was available from a variety of sources, including death certificates. Our parameter of interest is the risk difference $\psi_0^F = P_{X,0}(Y_1 = 1) - P_{X,0}(Y_0 = 1)$, the average treatment effect of LTPA on mortality five years after baseline interview.

The cohort was reduced to a size of $n = 2066$, as 26 subjects were missing LTPA values and/or self-rated health score (1.2% missing data). The prevalence of death was 13%, and the number of cases in the cohort sample was $nC = 269$. The TMLE was estimated on the full cohort sample, and the results are displayed in Table 5.2. Within TMLE, the machine learning Deletion/Substitution/Addition (DSA) algorithm (Sinisi and van der Laan 2004) was used to obtain an estimate of the functions

$\bar{Q}_0 = P_{X,0}(Y = 1 \mid A, W)$ and $g_0 = P_{X,0}(A \mid W)$ since the functional form of the data was unknown. Alternatively, one could also use an ensemble approach, such as super learning. The estimated parameter of interest was highly significant, and indicates that physical activity at or above recommended levels decreases five-year mortality risk in this population by 5.4%.

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

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

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

Table 5.1: SPPARCS variables

Variable	Description
Y	Death occurring within 5 years of baseline
A	LTPA score $\geq$ 22.5 METs at baseline
W	Health self-rated as "excellent" Health self-rated as "fair" Health self-rated as "poor" Current smoker Former smoker Cardiac event prior to baseline Chronic health condition at baseline $x \leq 60$ years old $60 < x \leq 70$ years old $80 < x \leq 90$ years old $x > 90$ years old Female

Table 5.2: SPPARCS cohort results. *The TMLE was estimated in the SPPARCS cohort. Sample size was 2066, with 269 deaths five years from baseline interview and 1797 nondeaths. RD is risk difference, SE is standard error, and p is p-value*

	Estimate	SE	p
RD	-0.054	0.012	< 0.001

Table 5.3: SPPARCS simulated nested case-control results. *IPCW-TMLEs were estimated in the nested case-control samples, and TMLEs were estimated in the cohort samples. RD is risk difference, SE is standard error, RE is relative efficiency compared to cohort RD, nC* $= 269$ *is number of cases, and nCo is number of controls*

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

in each of the 1000 nested case-control samples, and the TMLE was estimated in the cohort samples. The DSA algorithm was used to obtain estimates of the functions $\bar{Q}_0$ and g_0. The relative efficiency of the nested case-control parameters are compared to the cohort parameter in Table 5.3, as well as average values for the parameter of interest. Relative efficiency of the nested case-control design improved as the number of controls increases. With an average of 4 controls per case (approximately 1076 of the 1797 available noncase subjects), the relative efficiency of the nested case-control design reached 78.9%.

Simulated Cohort

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

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

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

Table 5.4: Simulation data nested case-control results. *IPCW-TMLEs were estimated in the nested case-control samples and TMLEs were estimated in the cohort samples. RD is risk difference, SE is standard error, RE is relative efficiency compared to cohort RD, $nC = 296$ is number of cases, and nCo is number of controls*

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

Table 5.5: Randomized controlled trial simulation data nested case-control results. *IPCW-TMLEs were estimated in the nested case-control samples and TMLEs were estimated in the full trial samples. SE is standard error, RE is relative efficiency compared to cohort RD, $nC = 647$ is number of cases, and nCo is number of controls*

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

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

The true value for the risk difference was $RD = -0.061$ and the prevalence of death was 13.3%. One cohort sample was taken with 2,066 individuals, and the estimated value of death prevalence was 14.3%. The number of cases in the cohort sample was $nC = 296$. Controls were randomly sampled from among the noncases in the original cohort at various sample sizes relative to the number of cases ($2nC$, $3nC$, and $4nC$), and assigned the value $\Delta = 1$. Noncases that were not sampled were assigned the value $\Delta = 0$. The values for $P_{X,n}(\Delta = 1 \mid Y = 0)$ were 0.330, 0.506, and 0.674 for the three sample sizes. All cases were assigned $\Delta = 1$. Logistic regression was used to estimate the functions $\bar{Q}_0$ and g_0 since the functional form was known.

The relative efficiency of the nested case-control parameters are compared to the cohort in Table 5.4, as well as average values for the parameter of interest. As before, relative efficiency of the nested case-control design improves as the number of controls increases. With an average of 4 controls per case, the nested design reaches a relative efficiency of 78.4%.

Simulated Clinical Trial

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

We proposed that the effect modifier of interest, $W_3 \equiv A^*$ was only measured in a nested case-control sample. Controls were randomly sampled from among the noncases in the original cohort at various sample sizes relative to the number of cases ($2nC$, $3nC$, $4nC$, and $5nC$), and assigned $\Delta = 1$. Noncases that were not sampled were assigned $\Delta = 0$. The values for $P_{X,n}(\Delta = 1 \mid Y = 0)$ were 0.141, 0.210, 0.280, and 0.350 for the four sample sizes. All subjects with $Y = 1$ were assigned $\Delta = 1$.

An IPCW-TMLE was used to analyze the nested case-control samples. Multinomial regression was used with main terms to estimate the function $\bar{Q}_0$, representing a misspecified model. Due to the double robustness of the TMLE and IPCW-TMLE procedures, the estimates of the parameter of interest are consistent even when $\bar{Q}_0$ is misspecified. The values for $g_0(A^* \mid W)$ were known since it was a randomized controlled trial. Results are displayed in Table 5.5. The relative efficiency of the nested case-control design improves as the number of controls increases, and with 38.8% of the total trial participants we reach an efficiency of 86.4%.

5.1.4 Discussion

Two-stage sampling designs, including nested case-control sampling, are popular in many fields, including epidemiology. They have the potential to reduce the costs associated with collecting data on the full cohort with minimal losses in efficiency (Ernster 1994; Rothman and Greenland 1998; Hak et al. 2004; Vittinghoff and Bauer 2006), as discussed in Chapter 1. We introduced the IPCW-TMLE for estimation of causal effects in two-stage sampling designs, with a focus on nested case-control sampling designs. In general, TMLE methodology can be used in conjunction with procedures that handle censoring, missingness, measurement error, and other persistent issues found in public health and medicine, in addition to adjusting for the missingness due to the two-stage sampling design.

Our simulated nested case-control studies within the SPPARCS data demonstrated 78.9% efficiency with an average of 4 controls per case. We had 78.4% efficiency in our simulated nested case-control studies within a simulated cohort, again with an average of 4 controls per case. These results coincided with the conclusions of Ury (1975), which noted that as a general rule, 4 controls per case yields a relative efficiency of 80.0%. We also demonstrated the use of IPCW-TMLEs for nested case-control study designs within randomized controlled trials when interested in an effect modification research question. With less than 40% of the trial subjects, we reached an efficiency of 86.4% compared to the full trial.

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

5.2 Prediction

Risk scores are calculated to identify those patients at the highest level of risk for an outcome. In some cases, interventions are implemented for patients at high risk. In population-based studies of the comparative effectiveness of treatments, patients may be matched or stratified based on their predicted risk of a disease or death. Standard practice for risk score prediction relies heavily on parametric regression. Generating a good estimator of the function of interest using parametric regression can be a significant challenge. High-dimensional data are increasingly common in epidemiology, and researchers may have dozens, hundreds, or thousands of potential predictors that are possibly related to the outcome. The complexity of the parametric regression may increase to the point that there are more unknown parameters than observations. Also, the best estimator of the true functional may be described by a complicated function not easily approximated by main terms or interaction terms.

The analysis of full cohort data for risk prediction is frequently not feasible, often due to the cost associated with purchasing access to large comprehensive databases, storage and memory limitations in computer hardware, or other practical considerations. Thus, researchers frequently conduct nested case-control studies instead of analyzing the full cohort, particularly when their prediction research question involves a rare outcome. This type of two-stage design introduces bias since the proportion of cases in the sample is not the same as the population. This complication may have contributed to the relative lack of prediction studies for rare diseases.

As previously discussed, an existing method for prediction in parametric statistical models with nested case-control samples is intercept adjustment. Traditional risk score approaches for prediction (e.g., logistic regression in a parametric statistical model) are not effective when based on case-control study data since the study design produces a biased sample. This complication may have contributed to the relative lack of prediction studies for rare diseases. Many published findings for prediction in rare diseases are based on the stratification of case-control samples (Whiteman and Green 2005; van der Steeg et al. 2007).

We consider a two-stage sampling design in which one takes a random sample from a target population and measures Y, the outcome, on each subject in the first stage. The second stage involves drawing a subsample from the original sample, collecting additional data on the subsample. The decision regarding selection into the subsample is influenced by Y. This data structure can be viewed as a missing-data structure on the full data structure X collected in the second stage of the study. Using nested case-control data from a Kaiser Permanente database, we generate a function for mortality risk score prediction using super learner and inverse-probability-of-missingness weights to correct the bias introduced by the sampling design.

5.2.1 Data, Model, and Parameter

Kaiser Permanente Northern California provided medical services to approximately 3 million members during the study period. They served 345,191 persons over the age of 65 in the 2003 calendar year, and 13,506 of these subjects died the subsequent year. The death outcome was ascertained from California death certificate filings. Disease and diagnosis variables, which we refer to in this paper simply as medical flags, were obtained from Kaiser Permanente clinical and claims databases. There are 184 medical flags covering a variety of diseases, treatments, conditions, and other reasons for visits. Gender and age variables were obtained from Kaiser Permanente administrative databases.

A nested case-control sample was extracted from the Kaiser Permanente database for computational ease. All 13,506 cases from the 2003–2004 data were sampled with probability 1, and an equal number of controls were sampled from the full database with probability 0.041 for a total of 27,012 subjects. Approval from the institutional review board at Kaiser Permanente Northern California for the protection of human subjects was obtained.

Formally, we define the full data structure as $X = (W, Y) \sim P_{X,0}$, with covariate vector $W = \{W_1, \dots W_{186}\}$ and binary outcome Y, indicating death in 2004. The observed data structure for a randomly sampled subject is $O = (Y, \Delta, \Delta X) \sim P_0$, where Y is included in X and Δ denotes the indicator of inclusion in the second-stage sample (nested case-control sample). The parameter of the full-data distribution of X is given by $\bar{Q}_0 = E_{X,0}(Y \mid W)$ and the full-data statistical model $\mathcal{M}^F$ is nonparametric.

5.2.2 Loss Function

Had our sample been comprised of n i.i.d. observations X_i, we would have estimated $\bar{Q}_0 = E_{X,0}(Y \mid W)$ with loss-based learning using loss function $L^F(X, \bar{Q})$. Given the actual observed data, we can estimate $\bar{Q}_0$ with super learning and weights $\Delta_i / P_{X,n}(\Delta_i = 1 \mid Y_i)$ for observations $i = 1, \dots, n$, which corresponds with the same super learner, but now based on the inverse-probability-of-missingness(/censoring)-

weighted loss function:

$$L(O, \bar{Q}) = \frac{\Delta}{P_{X,n}(\Delta = 1 \mid Y)} L^F(X, \bar{Q}).$$

We define our parameter of interest as: $\bar{Q}_0 = \arg\min_{\bar{Q}} E_0 L(O, \bar{Q})$, where $\bar{Q}$ is a possible function in the parameter space of functions that map an input W into a predicted value for Y. $E_0 L(O, \bar{Q})$, the expected loss, evaluates the candidate $\bar{Q}$, and it is minimized at the optimal choice of $\bar{Q}_0$.

5.2.3 Data Analysis

We implemented super learning with observation weighting in R (R Development Core Team 2010) to obtain our estimate of $\bar{Q}_0$ using our observed data. Using a server with dual quad-core Intel E5420 processors running at 2.50GHz and 64GB of memory, our analysis using the SuperLearner and CV.SuperLearner functions from the SuperLearner package in R took 21 hours. CV.SuperLearner calculated the cross-validated risk for the super learner algorithm.

Observation weights within the super learner were assigned based on the inverse probability of missingness, $w_i = \Delta_i / P_{X,n}(\Delta_i = 1 \mid Y_i)$ thus cases were given observation weights equal to 1 and controls were given observation weights of $1/0.041 = 24$. One could further stabilize the weights by standardizing them to sum to 1: in other words, we would divide the above w_i by $\sum_{i=1}^{n} \Delta_i / P_{X,n}(\Delta = 1 \mid Y_i)$. Recall that the super learner allows a researcher to use multiple algorithms to outperform a single algorithm in nonparametric and semiparametric statistical models by taking a weighted average of the algorithms. Any algorithm that allows observation weighting can be used with super learner in nested case-control data.

The collection of 16 algorithms included in this analysis can be found in Table 5.6. We implemented dimension reduction among the covariates as part of each algorithm, retaining only those covariates associated with Y in a univariate regression ($p < 0.10$). After screening, 135 covariates remained. Algorithms with different options (e.g., degree, size, etc.) were considered distinct algorithms. The selection of these algorithms was based on investigator knowledge, the ability to take observation weights, and computational speed.

A summary of the nested case-control variables can be found in Table 5.7. All 187 variables, except death, were evaluated from 2003 records. The majority of the sample is female, with 45.2% male. The age category with the largest number of members was 70 to 79, with 41.0%. (For presentation, age is summarized categorically in Table 5.7, although the variable is continuous and was analyzed as a continuous variable. All other variables are binary.) The top ten most prevalent medical flags in the sample were: screening/observation/special exams, other endocrine/metabolic/nutritional, hypertension, minor symptoms, postsurgical status/aftercare, major symptoms, history of disease, other musculoskeletal/connective tissue, cataract, and other dermatological disorders. The majority of medical flags

Table 5.6: Collection of algorithms

Algorithm	Description
glm.1	Main terms logistic regression
glm.2	Main terms logistic regression with gender × age interaction
glm.3	Main terms logistic regression with gender × age^2 interaction
glm.4	Main terms logistic regression with gender × age^3 interaction
glm.5	Main terms logistic regression with age^2 term
glm.6	Main terms logistic regression with age^3 term
glm.7	Main terms logistic regression with age × covariate interaction for remaining main terms
glm.8	Main terms logistic regression with gender × covariate interaction for remaining main terms
glm.9	Main terms logistic regression with age × covariate and gender × covariate interaction
bayesglm	Bayesian main terms logistic regression
glmnet.1	Elastic net, $\alpha = 1.00$
glmnet.5	Elastic net, $\alpha = 0.50$
gam.2	Generalized additive regression, degree = 2
gam.3	Generalized additive regression, degree = 3
nnet.2	Neural network, size = 2
nnet.4	Neural network, size = 4

(47.2%) had a prevalence of less than 1%. Twenty medical flags had a prevalence of 0%. These variables were excluded from our analysis as they provide no information. We remind the reader that these percentages do not reflect estimates of prevalence in the *population* given the biased sampling design.

The super learning algorithm for predicting death (risk score) in the nested case-control sample performed as well as or outperformed all single algorithms in the collection of algorithms. With a cross-validated MSE (i.e., the cross-validated risk, not to be confused with *risk score*) of 3.336e-2, super learner improved upon the worst algorithms by 17% with respect to estimated cross-validated MSE. MSEs in the collection of algorithms ranged from 3.336e-2 to 3.913e-2. While the collection of algorithms was somewhat limited, which isn't optimal from a theoretical perspective, we see some benefits in relative efficiency. Results are presented in Table 5.8 where relative efficiency for each of the k algorithms is defined as RE=cross validated MSE(k)/cross validated MSE(*super learner*).

When examining R^2 values, the super learner had the largest R^2 compared to the collection of algorithms with an $R^2 = 0.113$, although ten of the algorithms approached this value. Super learner had an 11.3% gain relative to using the marginal probability (i.e., assigning probability of death 0.039 to each observation). The algorithms in the collection had R^2 values ranging from 0.112 to -0.041. (Negative

Table 5.7: Characteristics of Northern California Kaiser Permanente members aged 65 years and older in nested case-control sample, 2003

Variables	No.	%
Death (in 2004)	13,506	50.0
Male	12,213	45.2
Age, years[a]		
65 to <70	5,193	19.2
70 to <80	11,077	41.0
80 to <90	8,525	31.6
$\geq$ 90	2,217	8.2
Most prevalent medical flags	No.	%
Screening/observation/special exams	23,597	87.4
Other endocrine/metabolic/nutritional	10,633	39.4
Hypertension	10,612	39.3
Minor symptoms, signs, findings	9,748	36.1
Postsurgical status/aftercare	9,447	35.0
Major symptoms, abnormalities	8,251	30.5
History of disease	7,376	27.3
Other musculoskeletal/connective tissue	7,359	27.2
Cataract	5,976	22.1
Other dermatological disorders	5,692	21.1
Medical flag prevalence	No.	%
Zero	20	10.8
$0 < x < 1\%$	67	36.4
$1 \leq x < 10\%$	72	39.1
$\geq 10\%$	25	13.6

[a] Age is summarized categorically although the variable is continuous.

R^2 values indicate that the marginal prevalence probability is a better predictor of mortality than the algorithm. Values for R^2 can fall outside the range [0,1] when calculated in cross-validated data.) See Table 5.8. While the performance of the super learner improved upon the collection of algorithms with respect to R^2 values, it should be noted that the overall prediction power of this data set is somewhat limited with the best $R^2 = 0.113$.

5.2.4 Discussion

Alternatives to parametric approaches to risk score prediction include the flexible approach super learning that provides improved performance in realistic nonparametric and semiparametric statistical models for high dimensional data. The algorithm provides a system to combine many estimators into an improved estimator and returns a function we can use for prediction in new data sets. Cross-validation of the individual algorithms and the super learner prevents overfitting and the selection of a fit that is too biased. Our criterion for estimator selection is based on an a priori established benchmark (e.g., cross-validated MSE). Thus, researchers are not limited to logistic regression in misspecified statistical models for prediction in case-control study designs.

Super learning allows for the use of observation weighting in order to generate prediction functions with nested case-control data, as well as data from other two-

Table 5.8: Results from super learner analysis

Algorithm	CV MSE	RE	R^2
SuperLearner	3.336e-2	–	0.113
glm.1	3.350e-2	1.004	0.109
glm.2	3.350e-2	1.004	0.109
glm.3	3.349e-2	1.004	0.109
glm.4	3.348e-2	1.004	0.109
glm.5	3.348e-2	1.004	0.109
glm.6	3.348e-2	1.004	0.109
glm.7	3.458e-2	1.037	0.080
glm.8	3.443e-2	1.032	0.084
glm.9	3.533e-2	1.059	0.060
bayesglm	3.778e-2	1.132	-0.005
glmnet.1	3.337e-2	1.000	0.112
glmnet.5	3.336e-2	1.000	0.112
gam.2	3.349e-2	1.004	0.109
gam.3	3.349e-2	1.004	0.109
nnet.2	3.913e-2	1.173	-0.041
nnet.4	3.913e-2	1.173	-0.041

stage sampling designs, case-control designs, and general biased sampling designs. We introduced a more flexible method for prediction in two-stage nested case-control data. This method is an application of the general loss-based super learner, and the appropriate loss function is selected. It corresponds with an inverse-probability-of-missingness full-data loss function. The method involves observation weights $w_i = \Delta_i/P_n(\Delta_i = 1 \mid Y_i)$ to eliminate the bias of the sampling design, where these weights are determined by the inverse probability of missingness. For nested case-control studies, this is equalivalent to using case-control weights, with cases assigned the weight q_n (an estimate of q_0 obtained from the full cohort) and controls assigned a weight of $(1 - q_n)/J$, where J is the average number of controls per case. Thus the choice of loss function can also be presented as the case-control-weighted loss function presented in the preceding two chapters.

In our nested case-control Kaiser Permanente data, super learner performed as well as or outperformed all algorithms in the collection of algorithms. While the overall predictive power of this data set was limited ($R^2 = 0.113$), the utility of super learning is still apparent. In Rose and van der Laan (2011), larger improvements in cross-validated MSE were seen in other real data sets. The minimal improvement of the super learner in this analysis is not unexpected since the outcome is rare in the population of interest. This can be understood intuitively since any large improvement in predicting death by an algorithm among "case" subjects is averaged over the entire sample.

It is not possible to know with certainty a priori which single algorithm will perform the best in any given data set. Even when the result is a negligible improvement relative to the best algorithms in the collection, the super learner provides a tool for researchers to run many algorithms and return a prediction function with the best cross-validated MSE, avoiding the need to commit to a single algorithm.

For example, even in this analysis, had the logistic regression with main terms and age covariate and gender covariate interactions for each covariate (glm.9) been the a priori selected single algorithm, with $R^2 = 0.060$, its performance is poor compared to that of the super learner. Several other algorithms were considerably worse thanglm.9 and also could have been the single a priori selected algorithm. In other words, the use of the super learner prevents poor a priori algorithm choices.

One might counter that their procedure would involve implementing multiple algorithms, using cross-validation and selecting the one algorithm with the best cross-validated MSE. This procedure is itself a super learning algorithm, referred to as the discrete super learner. The discrete super learner algorithm must then also be cross-validated in order to assess its performance. Once a discrete super learner has been implemented, only the relatively trivial calculation of the weight vector needs to be completed to implement the super learner. Super learning is an effective method for prediction, but recall that it also has applications in effect estimation. As discussed in recent epidemiology articles (Sudat et al. 2010; Snowden et al. 2011; Rose et al. 2011), researchers are frequently concerned about parametric model misspecification within effect estimation procedures and may wish to implement methods such as super learning. Super learning can be applied to a broad range of applied problems in epidemiology and medicine. Risk score prediction, designs based on propensity score matching, and incorporation into effect estimation procedures are just a few of these areas. The study presented in this paper further demonstrates the promise of the super learner illustrated in previous publications (van der Laan et al. 2007; Polley and van der Laan 2009).

Appendix: Wang et al. and IPCW-TMLE

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

If $X = (S, Y, A, W)$, and one only assumes the consistency and randomzation assumption, then the statistical model for the distribution of X is indeed nonparametric. Thus, in that statistical model, the proposed IPCW-TMLE of $EY(a)$ will be efficient if S, Y, A are discrete and $P_{X,0}(\Delta = 1 \mid S, Y, A)$ is estimated nonparametrically or in targeted manner. However, as in Wang et al., if one also assumes a parametric model for the treatment mechanism, then the statistical model for the full-data is *not* nonparametric. As a consequence of this choice of full-data model, the efficient influence curve does not exist in closed form, and has smaller variance than the efficient influence curve for the nonparametric full-data model, (and there exists a whole class of double robust influence curves/estimating functions), so that the Cramer-Rao lower bound in their more restricted model is smaller than the Cramer-Rao lower bound for the nonparametric full-data model our IPCW-TMLE aims to achieve. For such a nonparametric full-data model, their locally efficient estimator solves the actual efficient influence curve estimating equation while the IPCW-TMLE solves the inefficient IPCW-full-data efficient equation.

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

Chapter 6

Concluding Remarks

Causal inference methods and nonparametric and semiparametric estimators for case-control studies had been previously underdeveloped in the literature. Given the popularity of these designs in the public health and medical literature, the impact of CCW-TMLE, the method proposed in this thesis, is significant. Case-control studies are attractive to investigators researching rare diseases where they are able to sample known cases instead of following a large number of subjects and waiting for disease onset among only a few individuals. Application areas include general epidemiology, many branches of medicine including cancer, and genomics.

The dearth of methodology for case-control studies may be due to the data-generating experiment involving an additional complexity called biased sampling. That is, one assumes the underlying experiment that randomly samples a unit from a target population, measures baseline characteristics, assigns an exposure, and measures a final binary outcome, but we sample from the conditional probability distribution, given the value of the binary outcome. And yet, we still desire to assess the causal effect of exposure on the binary outcome for the target population.

After presenting a thorough literature review of related existing methodology in Chapter 1, we presented the targeted learning framework in Chapter 2. The CCW-TMLE is presented in Chapter 3, and it relies on knowledge of the true prevalence probability, or a reasonable estimate of this probability, in case-control weights to eliminate the bias of the sampling design. This case-control weighting scheme maps the TMLE for a random sample into a method for case-control sampling. Our simulation studies demonstrated that an existing method, the "approximately correct" IPTW estimator, has greater bias and is less efficient than the CCW-TMLE, in some cases, to an extreme degree. We also presented the CCW-MLE, which also had improved performance over the IPTW estimator.

Individual matching in case-control studies has, at times, been implemented possibly due to a misunderstanding of the true benefits of such a design. These designs are quite common, and while matching is intended to eliminate confounding, the main *potential* benefit of matching in case-control studies is a gain in efficiency. In Chapter 4, we presented and investigated the use of CCW-TMLE in matched case-control study designs. We also compare the CCW-TMLE in matched and unmatched designs in an effort to determine which design yields the most information about the causal effect. Our simulations supported the literature: in many practical situations researchers may be better served using an unmatched design.

Lastly, in Chapter 5, we considered two-stage sampling designs, including so-called nested case-control studies, where one takes a random sample from a target population and completes measurements on each subject in the first stage. The second stage involves drawing a subsample from the original sample, collecting additional data on the subsample. This data structure is truly a missing data structure on the full-data structure collected in the second stage of the study. We proposed an IPCW-TMLE and an inverse-probability-of-censoring-weighted super learner for two-stage sampling designs. Our IPCW-TMLE simulations demonstrated that one can achieve nearly 80% efficiency (compared to an analysis of the full data) in observational data using a two-stage design with an average of 4 controls per case. We also presented an analysis using super learner in nested case-control data from a large Kaiser Permanente database to generate a function for mortality risk prediction where the inverse-probability-of-censoring-weighted super learner performed as well as or better than the candidates included in collection of algorithms.

> The road map for targeted learning provides a recipe for researchers to investigate parameters they truly care about under realistic assumptions using various study designs, including case-control study designs.

References

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

K.M. Anderson, P.W.F. Wilson, P.M. Odell, and W.B. Kannel. An updated coronary risk profile. a statement for health professionals. *Circulation*, 83:356–362, 1991.

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

W.E. Barlow, E. White, R. Ballard-Barbash, P.M. Vacek, L. Titus-Ernstoff, P.A. Carney, J.A. Tice, D.S. Buist, B.M. Geller, R. Rosenberg, B.C. Yankaskas, and K. Kerlikowske. Prospective breats cancer risk prediction model for women undergoing screening mammography. *J Natl Cancer Inst*, 98(17):1204–1214, 2006.

O. Bembom and M.J. van der Laan. A practical illustration of the importance of realistic individualized treatment rules in causal inference. *Electron J Stat*, 1: 574–596, 2007.

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. Bembom, M.L. Petersen, 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. *Stat Med*, 28:152–72, 2009.

W.Z. Billewicz. The efficiency of matched samples: an empirical investigation. *Biometrics*, 21(3):623–644, 1965.

L. Breiman. Stacked regressions. *Mach Learn*, 24:49–64, 1996.

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

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

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.

N.E. Breslow, N.E. Day, K.T. Halvorsen, R.L. Prentice, and C. Sabal. Estimation of multiple relative risk functions in matched case-control studies. *Am J Epidemiol*, 108(4):299–307, 1978.

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

F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation and sparsity via L1 penalized least squares. In G. Lugosi and H.-U. Simon, editors, *COLT*, volume 4005 of *Lecture Notes in Computer Science*, Berlin Heidelberg New York, 2006. Springer.

F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation for gaussian regression. *Ann Stat*, 35(4):1674–1697, 2007a.

F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Sparse density estimation with L1 penalties. In N.H. Bshouty and C. Gentile, editors, *COLT*, volume 4539 of *Lecture Notes in Computer Science*, Berlin Heidelberg New York, 2007b. Springer.

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

W.G. Cochran. Matching in analytical studies. *Am J Public Health*, 43:684–691, 1953.

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

W.G. Cochran. The planning of observational studies of human populations. *J R Stat Soc Ser A Gen*, 128(2):234–266, 1965.

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 3rd Berkeley symposium, Vol IV*, Berkeley, 1956. University of California Press.

J.P. Costantino, M.H. Gail, D. Pee, S. Anderson, C.K. Redmond, J. Benichou, and H.S. Wieand. Validation studies for models projecting the risk of invasive and total breast cancer incidence. *J Natl Cancer Inst*, 91(18):1541–1548, 1999.

M.C. Costanza. Matching. *Prev Med*, 24:425–433, 1995.

A.S. Dalalyan and A.B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In N.H. Bshouty and C. Gentile, editors, *COLT*, volume 4539 of *Lecture Notes in Computer Science*, Berlin Heidelberg New York, 2007. Springer.

A.S. Dalalyan and A.B. Tsybakov. Aggregation by exponential weighting, sharp pac-Bayesian bounds and sparsity. *Mach Learn*, 72(1–2):39–61, 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. *Am Heart J*, 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 Med Res Meth*, 5(5), 2005.

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

R. Freedman. Incomplete matching in ex post facto studies. *Am J of Soc*, 55(5): 485–487, 1950.

M.H. Gail, L.A. Brinton, D.P. Byar, D.K. Corle, S.B. Green, C. Schairer, and J.J. Mulvihill. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst*, 81(24): 1879–1886, 1989.

O. Gefeller, A. Pfahlberg, H. Brenner, and J. Windeler. An empirical investigation on matching in published case-control studies. *Eur J Epidemiol*, 14:321–325, 1998.

S. Geisser. The predictive sample reuse method with applications. *J Am Stat Assoc*, 70(350):320–328, 1975.

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

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. Gruber and M.J. van der Laan. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *Int J Biostat*, 6(1), 2010a.

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

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.

T.R. Holford, C. White, and J.L. Kelsey. Multivariate analysis for matched case-control studies. *Am J Epidemiol*, 107(3):245–255, 1978.

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

R. Jackson. Updated new zealand cardiovascular disease risk-benefit prediction guide. *Br Med J*, 320(7236):709–710, 2000.

A. Juditsky, A.V. Nazin, A.B. Tsybakov, and N. Vayatis. Generalization error bounds for aggregation by mirror descent with averaging. In *NIPS*, 2005.

W.B. Kannel, D. McGee, and T. Gordon. A general cardiovascular risk profile: the Framingham study. *Am J Cardiol*, 38:46–51, 1976.

L.L. Kupper, A.J. McMichael, and R. Spirtas. A hybrid epidemiologic study design useful in estimating relative risk. *J Am Stat Assoc*, 70(351):524–528, 1975.

L.L. Kupper, J.M. Karon, D.G. Kleinbaum, H. Morgenstern, and D.K. Lewis. Matching in epidemiologic studies: validity and efficiency considerations. *Biometrics*, 37:271–291, 1981.

M. LeBlanc and R.J. Tibshirani. Combining estimates in regression and classification. *J Am Stat Assoc*, 91:1641–1650, 1996.

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 Gen*, 140:469–491, 1977.

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.

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

S.M. McKinlay. Pair-matching – a reappraisal of a popular technique. *Biometrics*, 33(4):725–735, 1977.

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

K.L. Moore and M.J. van der Laan. Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. *Stat Med*, 28(1): 39–64, 2009b.

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

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.

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

E.C. Polley and M.J. van der Laan. Predicting optimal treatment assignment based on prognostic factors in cancer patients. In K.E. Peace, editor, *Design, Summarization, Analysis & Interpretation of Clinical Trials with Time-to-Event Endpoints*, Boca Raton, 2009. Chapman & Hall.

E.C. Polley and M.J. van der Laan. Super learner in prediction. Technical Report 266, Division of Biostatistics, University of California, Berkeley, 2010.

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.

R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, 2010. URL `http://www.R-project.org`.

M. Rahman. Analysis of matched case-control data: Author reply. *J of Clin Epidemiol*, 56(8):814, 2003.

L.E. Ramsay, I.U. Haq, P.R. Jackson, and W.W. Yeo. Sheffield risk and treatment table for cholesterol lowering for primary prevention of coronary heart disease. *Lancet*, 346(8988):1467–1471, 1995.

L.E. Ramsay, I.U. Haq, P.R. Jackson, and W.W. Yeo. The Sheffield table for primary prevention of coronary heart disease: corrected. *Lancet*, 348(9036):1251, 1996.

J.M. Robins. [Choice as an alternative to control in observational studies]: Comment. *Stat Sci*, 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. *J Am Stat Assoc*, 89(427):846–866, 1994.

S. Rose and M.J. van der Laan. Simple optimal weighting of cases and controls in case-control studies. *Int J Biostat*, 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. *Int J Biostat*, 5(1):Article 1, 2009.

S. Rose and M.J. van der Laan. A targeted maximum likelihood estimator for two-stage designs. *Int J Biostat*, 7(1):Article 17, 2011.

S. Rose, J.M. Snowden, and K.M. Mortimer. Rose et al. respond to "G-computation and standardization in epidemiology". *Am J Epidemiol*, 173(7):743–744, 2011.

M. Rosenblum and M.J. van der Laan. Targeted maximum likelihood estimation of the parameter of a marginal structural model. *Int J Biostat*, 6(2):Article 19, 2010.

M. Rosenblum, S.G. Deeks, M.J. van der Laan, and D.R. Bangsberg. The risk of virologic failure decreases with duration of HIV suppression, at greater than 50% adherence to antiretroviral therapy. *PLoS ONE*, 4(9): e7196.doi:10.1371/journal.pone.0007196, 2009.

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

J.J. Schlesselman. *Case-Control Studies: Design, Conduct, Analysis.* Oxford, Oxford, 1982.

S.E. Sinisi and M.J. van der Laan. Deletion/Substitution/Addition algorithm in learning with applications in genomics. *Stat Appl Genet Mol*, 3(1), 2004. Article 18.

J.M. Snowden, S. Rose, and K.M. Mortimer. Implementation of g-computation on a simulated data set: demonstration of a causal inference technique. *Am J Epidemiol*, 173(7):731–738, 2011.

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

O.M. Stitelman and M.J. van der Laan. Targeted maximum likelihood estimation of time-to-event parameters with time-dependent covariates. Technical Report, Division of Biostatistics, University of California, Berkeley, 2011a.

O.M. Stitelman and M.J. van der Laan. Targeted maximum likelihood estimation of effect modification parameters in survival analysis. *Int J Biostat*, 7(1), 2011b.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *J R Stat Soc Ser B*, 36(2):111–147, 1974.

S.E. Sudat, E.J. Calton, E.Y. Seto, R.C. Spear, and A.E. Hubbard. Using variable importance measures from causal inference to rank risk factors of schistosomiasis infection in a rural setting in china. *Epidemiol Perspect Innov*, 7:3, 2010.

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

I. Tager, M. Hollenberg, and W. Satariano. Self-reported leisure-time physical activity and measures of cardiorespiratory fitness in an elderly population. *Am J Epidemiol*, 147:921–931, 1998.

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

A.B. Tsybakov. Optimal rates of aggregation. In B. Schölkopf and M.K. Warmuth, editors, *COLT*, volume 2777 of *Lecture Notes in Computer Science*, Berlin Heidelberg New York, 2003. Springer.

J. Tyrer, S.W. Duffy, and J. Cuzick. A breast cancer prediction model incorporating familial and personal risk factors. *Stat Med*, 23(7):1111–1130, 2004.

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. Estimation based on case-control designs with known prevalance probability. *Int J Biostat*, 4(1):Article 17, 2008a.

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

M.J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite sample oracle inequalities and examples. Technical Report 130, Division of Biostatistics, University of California, Berkeley, 2003.

M.J. van der Laan and S. Gruber. Collaborative double robust penalized targeted maximum likelihood estimation. *Int J Biostat*, 6(1):Article 17, 2010.

M.J. van der Laan and J.M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer, Berlin Heidelberg New York, 2003.

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

M.J. van der Laan and Daniel B. Rubin. Targeted maximum likelihood learning. *Int J Biostat*, 2(1):Article 11, 2006.

M.J. van der Laan, E.C. Polley, and A.E. Hubbard. Super learner. *Stat Appl Genet Mol*, 6(1):Article 25, 2007.

W.A. van der Steeg, S.M. Boekholdt, E.A. Stein, K. El-Harchaoui, E.S. Stroes, M.S. Sandhu, N.J. Wareham, J.W. Jukema, R. Luben, A.H. Zwinderman, J.J.P. Kastelein, and K.-T. Khaw. Role of the apolipoprotein B-apolipoprotein A-I ratio in cardiovascular risk assessment: a case-control analysis in EPIC-Norfolk. *Ann Intern Med*, 146:640–648, 2007.

J.P. Vandenbrouke, E. von Elm, D.G. Altman, P.C. Gotzsche, C.D. Mulrow, S.J. Pocock, C. Poole, J.J. Schlesselman, and M. Egger for the STROBE Initiative. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. *PLoS Med*, 4(10):1628–1654, 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.

H. Wang, S. Rose, and M.J. van der Laan. Finding quantitative trait loci genes with collaborative targeted maximum likelihood learning. *Stat Prob Lett*, published online 11 Nov (doi: 10.1016/j.spl.2010.11.001), 2010.

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

D.C. Whiteman and A.C. Green. A risk prediction tool for melanoma? *Cancer Epidemiol Biomarkers Prev*, 14(4):761–763, 2005.

P.W.F. Wilson, R.B. D'Agostino, D. Levy, A.M. Belanger, H. Silbershatz, and W.B. Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97:1837–1847, 1998.

D. H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.