Title
Applications of Semi-parametric Estimation Methods in Causal Inference and Prediction

Author
Jamshidian, Farid

Publication Date
2011

Peer reviewed|Thesis/dissertation

Applications of Semi-parametric Estimation Methods in Causal Inference and Prediction

by

Farid Jamshidian

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 Nicholas Jewell, Chair
Professor Alan Hubbard
Professor John Colford

Fall 2011

Abstract

Applications of Semi-parametric Estimation Methods

in Causal Inference and Prediction

by

Farid Jamshidian

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Nicholas Jewell, Chair

In this thesis, we argue for the use of loss-based semi-parametric estimation methods as an alternative to traditional parametric models in causal inference and prediction. We present a brief discussion on "black box" epidemiology in the first chapter and argue that risk factor epidemiology can be improved by using semi-parametric estimation methods. We demonstrate the use of semi-parametric methods by applying them to two different problems: one in causal inference and another in prediction. In each case, we demonstrate the process one would follow to define the question of interest, parameterize this question, and estimate it using semi-parametric methods.

In the second chapter we introduce a formal concept of a perception effect, and define unmasking and placebo effects in the context of randomized trials. We employ modern tools from causal inference to derive semi-parametric estimators of such effects. The methods are illustrated on a motivating example from a recent pain trial where the occurrence of treatment-related side effects acts as a proxy for unmasking.

In the third chapter, we redefine perception and unmasking effects for a longitudinal setting, and explore various causal graphs for the gabapentin trial. We demonstrate application of the semi-parametric methods in this more general setting by assuming a more complicated causal graph. To estimate the parameters, we use Maximum Likelihood Estimation and two different versions of Targeted Maximum Likelihood Estimation.

Finally, in chapter four, we approach coronary heart disease risk prediction modeling from a semi-parametric perspective using data from the Framingham study. The "super learner" is used with a library of machine learning algorithms to create an ensemble risk prediction model for coronary heart disease. We define relative risk importance parameters for various risk factors and estimate them with semi-parametric methods used in earlier chapters. The results are compared to the Framingham study and those obtained by fitting a parametric model to the Framingham dataset.

To Mehdi, Mina, Mortaza, Elvira, and Hossein.

"I would rather discover one causal law than be king of Persia."

Democritus (460-370 B.C.)

Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisors Prof. Nicholas Jewell and Prof. Alan Hubbard for their continuous support of my research, their patience, motivation, and immense knowledge. Their guidance helped me throughout my research and writing of this thesis. I could not have imagined having better mentors for my PhD than Nick and Alan.

In addition to my advisors, I would like to thank my thesis committee member, Prof. John Colford for all the time he spent on this thesis and for his very insightful comments.

My sincere thanks also goes to Dr. Steve Selvin, Dr. Mark van der Laan, Dr. Sandrine Dudoit, and Dr. Maureen Lahiff for everything they taught me during my years at Berkeley.

Last but not the least, I would like to thank my family; my father Mehdi for inspiring me to further my education, my mother Mina for her unconditional love and support, my uncle Mortaza for introducing me to statistics, and my brother Hossein for allowing me to be a child again. I would specially like to thank my fiancé Elvira for her endless hours spent editing and proofreading.

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

During the past few decades, risk factor epidemiology ("black box" epidemiology), vaguely defined as speculative linking of exposures and a diseases in hope of discovering the causes of the disease[1,2], has been the method of choice in most epidemiological studies. Risk factor epidemiology has been described by some as the "unique virtue" of the field[3], and by others as an "embarrassing liability".[1] Weed[4] refers to the discussion in the literature on the risk factor approach as the "black box" debate and traces it back to Peto.[4,5] In his research on cancer epidemiology, Peto described two different approaches: the first approach emphasized the underlying biological mechanism while the second approach, "the black box strategy", ignored the biology in favor of behavioral risk associations.[5] Even though the definition of black box epidemiology has been primarily used for explanatory studies, it can also be extended to encompass hypothetical linking of exposures to a disease for the purpose of prediction. Such linkage is often motivated by black box association studies. For instance, Gail et al.[6] incorporated age at first live birth into a risk prediction model for breast cancer based on previously established association between the exposure and the outcome.

Proponents of black box epidemiology have argued that it allows disease prevention in the absence of a clear understanding of the disease mechanism,[2,3]and that it has occasionally identified useful potential interventions.[2] Savitz argues that well designed epidemiological studies carefully evaluate the data observed in studies outside epidemiology, and suggests that legitimate epidemiological studies contribute to science and public health whether "we ignore, build upon, or contradict parallel information derived from other disciplines".[3] Greenland further defends black box epidemiology for its hypothesis generating ability, as a "valuable source of seemingly unrelated facts that await coherent explanation by novel theories".[2] According to Greenland, the purely descriptive approaches (publication of observations) to epidemiological studies are beneficial since such observations supply data for the scientific community to test their theories.[2] Yet, Greenland recognizes that a legitimate problem of risk-factor epidemiology is "over-interpretation of observed associations as causal".[2]

On the other hand, opponents of risk factor epidemiology have criticized the method for ignoring the underlying theory.[1] They argue that risk factor epidemiology lacks an underlying biological hypothesis, and thus, the method is no more than "stabs in the dark" by which researchers randomly link various exposures to various diseases.[1,3] Critics further argue that black box epidemiology has produced too many false alarms and go as far as describing much of epidemiology as wasteful research. [1,2] Skrabanek rejects Savitz's argument for contribution of black box studies to science by asserting that risk factor epidemiology is an ancillary methodology.[1] Skrabanek

argues that although risk factor epidemiology may provide testable hypotheses of causality if governed by scientific principles, it cannot contradict valid scientific data.[1] Overall, risk factor epidemiology has been disparaged for its failure to account for various biases; from inadequate adjustment for confounding, to biases due to the methods used for linking the exposure and the outcome and inference regarding the association.

The driving force behind the black box debate has been epidemiology's search for a new identity; a search for regularities. As others have suggested before,[2,7] we believe that regularities can be established without reliance on explanatory theories for the underlying mechanisms. Greenland asserts that one definition of "black box" research is searching for statistical regularities (data mining).[8] He notes the advancements of "black box" statistics in machine learning when the goal is purely prediction.[8] We believe that reducing explanatory problems in epidemiology to statistical prediction problems borrows from regularities already established in fields such as machine learning.

The black box does not only contain the underlying biological mechanisms that are ignored by the risk factor epidemiologists, but it also includes various statistical modeling assumptions and other biases that find their way into the box. Even if statistical models are used for purely "descriptive purposes", the interpretations of such descriptive statistics may not be beneficial to science as Greenland argues. Many false hypotheses may be generated based on such descriptive statistics and much valuable resources may be wasted to falsify them. Relying on the underlying biological theory may not be necessary for linking an exposure to an outcome; however, all other biases need to be accounted for, including biases due to model mis-specification. Parametric models are almost always misspecified. An alternative approach to parametric modeling is to use semi-parametric models that do not model redundant components. Using semi-parametric models can benefit risk factor epidemiology since such models lack biases due to misspecifications of the functional form of a parametric model. Semi-parametric methods accomplish the goal of reducing explanatory problems to prediction problems by using predictive tools for estimation of a particular parameter of interest. However, such statistical methods are not a substitute for a poor design, and our argument only applies to well designed studies with a well defined question of interest.

In the next few chapters, we demonstrate the use of semi-parametric methods in two different areas: explanation and prediction. In the second chapter, we use semi-parametric efficient estimation methods to estimate treatment and perception effects in a randomized clinical trial. We demonstrate the use of these methods in a more general longitudinal setting in chapter 3. Finally, in chapter 4, we apply loss-based semi-parametric estimation methods to the Framingham study to predict risk of coronary heart disease.

CHAPTER 2

Perception, Placebo, and Unmasking Effects in Randomized Clinical Trials

2.1 Background

Masking of participants and investigators has long been used in randomized clinical trials to prevent the measurement of research outcomes from being influenced by either the placebo effect or observer bias associated with knowledge of treatment assignment. This is of particular importance in clinical trials with a subjective patient response, or when the assessment of the outcome is quantified by clinicians. During the course of blinded clinical trials some patients, however, might be inadvertently unmasked to their assigned treatment, or at least grow to believe they are in a specific arm for several reasons. For instance, patients who receive treatment may believe they are on treatment as a result of experiencing documented treatment-related side effects, and/or placebo patients may believe that they are on placebo due to a lack of efficacy or worsening of their condition. Such "unmasking" may subsequently affect outcome reporting. In these clinical trials, the investigator may be interested in the causal effect of treatment, had unmasking not occurred with the patient's perception of treatment assignment remaining at a fixed baseline level.
In this chapter, we use recent methodology from the causal inference literature to formally define and estimate perception, placebo, and masking effects as theoretical interventions in a graphical model. We define causal treatment effects (after removal of unmasking effects) in terms of Type I and Type II direct effects based on a counterfactual framework[9], and estimate these effects using two semi-parametric estimation methods, Maximum Likelihood Estimation(MLE) and Targeted Maximum Likelihood Estimation (TMLE). To motivate our discussion, these definitions and estimation methods will be applied to data from a recent clinical trial that was conducted to study the effect of gabapentin for treatment of painful neuropathy among diabetic patients.

One of the most discomforting symptoms among diabetic patients is pain associated with peripheral neuropathy, estimated to affect about 45 percent of diabetic patients.[2] Backonja et al.[11] published results of a randomized double-blind clinical trial, conducted to evaluate the effect of a drug called gabapentin (or Neurontin) on pain among patients with either type I or type II diabetes. The study consisted of a seven-day screening phase, followed by an eight-week double-blind phase, and 165 participants were randomized to treatment or placebo groups. Baseline covariates such as age, height, weight, race, sex, and baseline pain and sleep scores were measured during the screening phase, prior to randomization. Treatment dosage was gradually titrated to a maximum tolerated dosage during the first four weeks of the double-blind phase, and it remained fixed thereafter. The main outcome measure was

daily pain severity, recorded by patients on an 11-point Likert scale (0-10) in daily diaries, and the primary endpoint was calculated as the mean score for the last seven recorded diary entries.[11]

Using an intent-to-treat (ITT) analysis, the investigators had reported that patients who received gabapentin had a significantly lower (p-value < 0.001) mean endpoint daily pain score than patients who received placebo with the treatment difference estimated to be a decline of 1.2 points. In what follows, we reanalyze the gabapentin trial data considering patients' perceptions and unmasking. For all our analysis we will use data on 164 participants (83 and 81 in the gabapentin and placebo groups, respectively) as one individual had reported no pain scores.

2.2 Non-Specific Effects of Treatment and the Placebo Effect

Although our focus is on the effects of perception on estimation of a causal treatment effect, it is worth beginning with a precise definition and brief discussion of the so-called placebo effect. Beecher first quantified this phenomenon in 1955. He observed that, in 15 trials studying different diseases, 35% of all 1082 patients were satisfactorily relieved by a placebo.[12] Many researchers have subsequently studied the placebo effect and have claimed significant improvements in patients' outcome due to this effect. However, a recent meta-analysis of clinical trials, including placebo and no-treatment groups, has questioned the significance of the placebo effect. Hrobjartsson and Gotzsche[13] performed a meta-analysis of 114 randomized trials on 40 different clinical conditions comparing treatments, placebo controls, and no treatment controls. Their goal was to investigate whether patients assigned to placebo had a better outcome than those assigned to no treatment. The study found no significant placebo effect on binary or continuous objective outcomes. The only consistent placebo effect was observed for continuous subjective outcomes. Among the 40 conditions, only trials with subjective pain score assessments as an outcome displayed a significant placebo effect across studies.

In earlier literature, the placebo effect is referred to a variety of responses that occur when patients are being treated with inactive placebo that in theory should have no therapeutic effect. This definition of a placebo effect is what we refer to as the placebo response (i.e. the outcome of a patient receiving a placebo), and is different than how a placebo effect is defined in more recent literature. Turner et al.[14] define the placebo effect as the non-specific effects of treatment attributable to factors other than specific active components. These non-specific effects include "physician attention, interest, and concern in a healing setting; patient and physician expectations of treatment effects; the reputation, expense, and impressiveness of the treatment; and characteristics of the setting that influence patients to report improvement." The latter definition may be thought of as the difference in a patient's outcome had he received a placebo compared to no treatment at all. These two differing definitions of the placebo effect have long been confused. Miller et al.[15] note that progress in understanding and estimating the placebo effect has been hampered by a lack of

conceptual clarity, some of which has been due to confusion of the placebo effect with the placebo response.

Consider the following hypothetical experiments: In a first experiment, a group of patients with a headache are observed without their knowledge. These patients receive neither a placebo nor any treatment. Any improvement in their condition must be solely due to the individual mechanisms of their bodies and/or interactions of the latter with a personalized environment. In a second experiment, a group of patients with a headache are given placebos by clinicians (with at least some expectation of a therapeutic effect as occurs when patients assume that there is some chance of receiving an active treatment), and their response is observed. Improvements in patients' condition for this experiment are due to such factors as: internal patient mechanisms, physician attention, patients' expectation regarding their assigned therapy, etc. The natural healing of the body cannot be attributed purely to the placebo effect as at least components of the effect are also present in the first experiment where patients do not receive a placebo. The placebo response is due to a combination of the placebo effect and internal patient mechanisms, and must be distinguished from the placebo effect. As Miller et al.[15] conclude, placebo-controlled trials are inadequate for elucidating the placebo effect, and to do so, we need no-treatment control groups.

Formally, we define the placebo effect for an individual as the difference in the outcome if an individual received a placebo as compared to no treatment at all. Although the presence of an active treatment does not directly factor into the definition, it is necessary for a placebo effect to exist. We return to this point further in the next section that discusses the impact of unmasking of treatment assignment. Quantification of specific and non-specific effects of treatment has received considerable attention in recent years. Petkova et al.[16] consider three hypothetical treatment scenarios of no treatment, placebo, and active treatment, combined with the counterfactual framework to separate specific and non-specific effects of treatment. To estimate the placebo effect, they compare the outcome of the placebo group to the baseline measurements of the outcome variable for all the patients, implicit from certain model assumptions.[16] Yet, this remains unsatisfactory in situations where measurements on patients at baseline may not be exchangeable for those arising later in a 'no treatment' group, as time itself is often an important factor influencing outcome assessment (as in the gabapentin trial).

2.3 Perception, Unmasking, and Side Effects as a Proxy

Masking patients in clinical trials prevents them from knowing certain information about the trial including the treatment group to which they are randomized. However, participants may develop a perception about their assigned treatment. Patients may either believe they are more likely to be on placebo, or more likely to be on treatment, or they may have no opinion about their treatment. In a more general sense, we may think of the patients assigning a degree of certainty to receiving a specific active

treatment. In a single treatment/placebo trial, a low degree of certainty would imply that the patient is leaning towards placebo and a high degree would mean that he thinks he is on treatment. We refer to this random variable of degree of certainty as a patient's *perception*, P, where $P = 1$ indicates that a patient is certain that he has been assigned the active treatment; at the other end of the scale, $P = 0$ indicates that a patient is certain that he is receiving the placebo.

In the extreme case where the investigator informs a patient about his treatment group, the patient would automatically have $P = 1$ or $P = 0$, depending on his original treatment assignment; such patients may be considered unmasked. We do not directly allow the variable P to distinguish here between the case where full unmasking has occurred and where an individual may be convinced that they are on active treatment (or placebo), even though this perception is incorrect. However, interaction effects between P and T allows differentiation of these two scenarios.

In almost all cases, we do not observe the patient's perception on a continuous scale, nor directly observe P. For simplicity, we focus on a discrete approximation. In particular, we consider observation of the following three-level variable,P, indicating perception, extending the simple version of P introduced above:

$$P = \begin{cases} 1 & patient\ believes\ she\ is\ on\ treatment \\ 0 & patient\ has\ no\ knowledge\ of\ assignment \\ -1 & patient\ believes\ she\ is\ on\ placebo \end{cases}$$

In double blind studies, experimenters are also masked (in addition to patients) to prevent patient outcomes from being influenced by the experimenter's expectations or interest. In such trials, data may be collected on experimenter's perception in a similar fashion. Even though our focus is on perception and unmasking of the patients, it is straightforward to expand our discussion to include perception/unmasking of investigators.

During the course of the gabapentin trial, some of the patients in both treatment and placebo arms developed a wide variety of side effects, many of which were known to be associated with active treatment. Backonja et al.[11] acknowledge that since the study end point is subjective, the occurrence of adverse events may result in unmasking of some patients, potentially, biasing the results of their efficacy analysis. The authors circumvented this problem by separately excluding patients with dizziness and somnolence, "the two most frequent adverse events, and also, the two with the largest difference in incidence between the treatment and the placebo groups".[11] After excluding patients with dizziness, the estimated mean endpoint pain score for the gabapentin group remained 1.2 (p-value = 0.002) points lower than the placebo group. Effectively, this stratifies participants by occurrence of this particular side effect and considers the results solely in the group who do not experience these adverse events. By a similar stratified analysis, when patients who reported somnolence were excluded, the treatment-placebo difference dropped to 0.81 points (p-value = 0.03). By analyzing side effects one-by-one, this approach does not address the simultaneous impact of all treatment-related side effects, nor does it

account for the effect of potential confounding variables which will be discussed in Section 6.

As in most clinical trials, patients in the gabapentin trial were not questioned on their perception regarding the treatment they received. It is likely, however, that the occurrence of any treatment-related side effects may have led patients to believe that they had been assigned to active treatment (incorrectly in some cases), the concern raised by the investigators leading to the naïve analysis above that excluded patients with a single such side effect. We thus use the occurrence of *any* treatment-related side effects as a proxy for perception P being set to $P = 1$. The list of treatment-related side effects includes amnesia, ataxia, depersonalization, insomnia, nervousness, etc. A total of 43 patients, 31 in the gabapentin group and 12 patients in the placebo group experienced at least one of these side-effects during the eight week period, the imbalance reflecting the anticipated association of these side effects with treatment. Patients who did not experience any side effects consisted of those who had no knowledge of the treatment assignment and those who believed they were receiving placebo. For our analysis, we label this group as $P = 0$, keeping in mind that it is a combination of the $P = -1$ and $P = 0$ groups introduced earlier. This point will be discussed further in the discussion section.

2.4 Type I and Type II Direct Effects

Consider three possible treatment "conditions" that may be assigned to all members of a population: (0) assigned to placebo, (1) assigned to active treatment, and (2) assigned to neither treatment nor placebo. Such assignments may not be ethical in some experiments with a major outcome, and we only consider them here as a hypothetical experiments. For each individual, define Y_j, to be the (possibly counterfactual) value of the outcome for individuals exposed to the j^{th} treatment $(j = 0,1,2)$. Then, the "causal" placebo effect could be defined as: $\psi_{placebo} = E[Y_0 - Y_2] = E[Y_0] - E[Y_2]$, that is, the population outcome mean when everyone receives the placebo in the experiment minus the outcome mean when all individuals receive neither treatment nor placebo. Note that it is important for the placebo effect to occur, that all individuals assigned to either treatment or placebo believe it is possible that they may receive active treatment. It is plausible and likely that the placebo effect, if it exists, depends on the perceived likelihood of receiving treatment. One may argue that the no-treatment control group is already unmasked by knowing they are not receiving a treatment, or any attempt to measure their outcome may affect the outcome itself, and therefore it is not possible to estimate the placebo effect. However, it may still be possible to estimate the placebo effect indirectly. For instance, consider cancer patients who visit their physician weekly to receive chemotherapy. At every visit, the patients are asked to rate their pain on a scale of 0 to 10 as part of a routine. The physician could (possibly) randomize some of these patients to a pain treatment group, some to a placebo group, and some to a no-treatment control group. The patients who are assigned to either treatment or placebo groups will be informed that they are part of a pain treatment study. The placebo

group may experience a placebo effect since they are aware of the possibility of being treated for pain, and they have a perceived likelihood of receiving the active treatment. Since the no-treatment group has no knowledge of the study or any expectations for improvement in their pain, and since their interaction with the clinicians is solely for cancer treatment, they may not experience a placebo effect. Typically, it is not feasible to observe a single individual in more than one experimental setting and so the chosen treatment $(j = 0,1,2)$ is randomly assigned to all individuals in the sample and in principle, this removes the potential for confounding, and allows population causal treatment and placebo effects to be estimated without bias. We now turn to the more common experiment where patients are solely allocated to either an active treatment or a placebo, so that the 'no treatment or placebo' group $(j = 2)$ is not evaluated.

For a variety of reasons individuals may vary on their perception, P, of their assigned treatment, as defined in its approximate form at the end of Section 3. Consider an ideal experiment in which the investigator measures the effect of a treatment A ($1 =$ treatment, $0 =$placebo) on the outcome holding all patients' perception at a fixed level $P = p$. In this ideal experiment, the direct effect of a treatment on an individual is defined as the difference in the counterfactual outcome if the individual received treatment with his perception fixed at level $P = p$ as compared to the counterfactual outcome if he received placebo with his perception again fixed at the same level $P = p$. Using standard notation, the Type I direct effect of a treatment on an individual can thus be written as $Y_{1,p} - Y_{0,p}$, where $Y_{a,p}$ denotes an individual's counterfactual outcome fixing both treatment and perception. The population direct effect of treatment at fixed perception level $P = p$ is thus $\psi_a(p) = E[Y_{1,p} - Y_{0,p}] = E[Y_{1,p}] - E[Y_{0,p}]$. Usually, in experiments where full blinding can be achieved and maintained, interest focuses on this mean with $P = 0$.

Although our focus will be on Type I direct effects in the rest of the chapter, it is worth mentioning the alternative Type II direct effects. In the simple example above, Type II direct effect could be defined as the difference in the counterfactual outcome if an individual received placebo as compared to his counterfactual outcome if he received treatment, with his perception held fixed at its counterfactual level under placebo. In the population, the mean Type II direct effect may thus be defined as $\psi_a(P_0) = E[Y_{1,P_0} - Y_{0,P_0}]$. In this case P_0 is the perception level of the individual (possibly counterfactual) if he received placebo.[9] Note that a type I direct effect of a treatment is defined at a fixed level of perception for everyone in the population. On the other hand, for a type II direct effect of a treatment, the perception levels may vary from one individual to another. In the definition of a type II direct effect, even though the perception levels vary among the patients, what patients have in common is that each individual's perception is fixed at what his perception would have been had he received a placebo (for more information on type II direct effects see Peterson et al.)[9,17].

One advantage of using a Type II direct effect is that it yields a single treatment/placebo comparison whereas with Type I direct effects, potentially different comparisons exist for each fixed level of *P*. For instance, considering a binary treatment $A = 0,1$, and a binary perception $P = 0,1$, there are two possible type I direct effects of the treatment at fixed levels of perception, namely, $\psi_a(1) = E[Y_{1,1} - Y_{0,1}]$, and $\psi_a(0) = E[Y_{1,0} - Y_{0,0}]$, whereas the type II direct effect is given by $\psi_a(P_0) = E[Y_{1,P_0} - Y_{0,P_0}]$. However, the definition of a Type II direct effect implied by a particular graph can be represented as a weighted average of the Type I direct effects across the different strata of P.[17] Thus, a Type II effect may obscure the variation of different Type I direct effects across the strata of perception levels.
In the following sections we discuss estimation of the Type I direct effects of treatment holding the post-randomization variable, perception, fixed. A general discussion of causal inference in the analysis of the impact of post-randomization factors can be found in Lynch et al.[18] A similar causal structure (see Section 6 below) in a quite different setting can be found in Rosenblum et al.[19] where the post-randomization factor is use of a secondary treatment or intervention that is not randomized.

2.5 Additional Parameters of Interest

In addition to treatment effects at various levels of fixed perception, we may also be interested in the perception effects at a particular fixed treatment level. For example, the perception effect for placebo patients may be defined as the difference in the average outcome had everyone received placebo and was masked (that is, $P = 0$) as compared to everyone receiving placebo but being unmasked (that is, $P = -1$). In general, we define perception effects based on two distinct perception levels p_i, p_j for $\neq j$. The perception effect on the outcome, due to having perception $P = p_i$ compared to $P = p_j$, at fixed treatment level a, may be defined as $\psi_p(p_i, p_j, a) = E[Y_{a,p_i}] - E[Y_{a,p_j}]$; thus $\psi_p(p_i, p_j, 0) = E[Y_{0,p_i}] - E[Y_{0,p_j}]$ for placebo patients, and $\psi_p(p_i, p_j, 1) = E[Y_{1,p_i}] - E[Y_{1,p_j}]$ for treated patients. This framework naturally lends itself to organizing the researchers' thoughts by defining the parameter of interest as the one they could have estimated when they could have performed any theoretical experiment of interest, and thus observed any set of counterfactuals of interest.

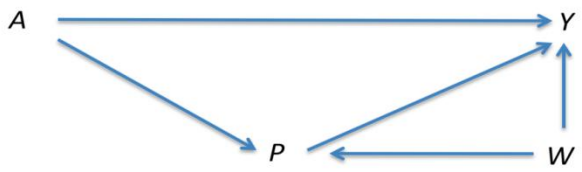
2.6 Estimation of Type I Direct Effects

The marginal effects of treatment and perception may be estimated using non-parametric or semi-parametric estimators (depending on the complexity of the involved covariates). Semi-parametric estimators for these marginal effects include: the Maximum Likelihood Estimator (MLE), the Inverse Probability of Treatment Weighted (IPTW) estimator[20,21], the Double Robust Inverse Probability of Treatment Weighted (DR-IPTW) estimator[22,23,24], and the Targeted Maximum Likelihood

Estimator (TMLE).[25,26] We focus on the MLE and TMLE, as the former is the simplest of the above estimators, and TMLE satisfies the properties of the most robust of these estimators (possessing the so-called double robust property discussed below). The TMLE also has the property that guarantees a proper model for the parameter of interest, as well as the observed data-generating distribution, which is not a guaranteed property of the estimating equation approaches (IPTW and DR-IPTW), and it is easy to implement in this case using standard software.

Let A denote treatment assignment, P denote perception, and Y be the outcome (as above). Also, let W represent a vector of baseline (pre-randomization) covariates. We have alluded to causal graph theory above, but now we discuss the consequences of assuming a particular Direct Acyclic Graph (DAG), one that describes one set of possible causal relationships between $A, P, W,$ and Y. A DAG is a directed graph formed by a collection of nodes (variables) and directed edges, each edge connecting one node to another. [27] The acyclic property of DAGs imposes a restriction on a directed graph such that no direct path can form a closed loop which starts from a node and returns to that node. [27]
Figure 2.1 illustrates a potential DAG that allows for (i) differing levels of perception across treatment groups, (ii) an effect of perception on the outcome, and (iii) a set of covariates that effect both perception and outcome. Note that since W is measured prior to treatment randomization, none of its components can lie on the causal pathway between A and Y. It is tempting to estimate the direct effects of treatment, controlling for perception, via simple stratification on P. However, Figure 2.1 and the rules developed for DAG's show that simply stratifying on perception, ignoring the covariates W, results in a biased estimate since this "introduces" confounding of the direct effect of treatment on the outcome. This arises since P is a collider in the DAG, being caused by both A and W. In graphical models, a variable on a path is called a collider if it is caused by two or more variables, since the arrows of the causing variables appear to collide on that node, blocking that path. [27]

Figure 2.1 A Direct Acyclic Graph linking treatment (A), perception (P), and covariates (W), to an outcome variable (Y).



Stratifying on colliders may result in new pathways being opened between the causing variables. Thus, if stratification/regression methods are to be used it is necessary to adjust for both P and W. For further discussion of these issues, see chapter 3 in Pearl[17] or chapter 9 in Jewell.[27] Conditioning on W raises difficulties if the set of covariates W is high-dimensional.

Returning to our general discussion, we note that the causal parameters defined in previous sections are defined in terms of all counterfactuals. However, we only observe one counterfactual for every subject. These parameters may still be estimated from observed data, however, if we make relevant assumptions. Different counterfactual outcomes are denoted as $Y_{a,p}$ for every value of $A = a$ and $P = p$, with the full data defined as $X^{FULL} = \left(W, Y_{a,p}, a \in A, p \in P\right)$, where we use A and P to denote the set of possible values of treatment and perception, respectively. First, we assume that the observed data for a subject can be treated as a random draw (according to some mechanism) of one of the counterfactuals from a theoretically defined full data, i.e. the observed data is assumed to be n i.i.d. copies of $O = (W, A, P, Y_{A,P})$,or a censored version of the theoretical full data consisting of all possible counterfactuals, X^{FULL}, (the so-called consistency assumption, closely related to the Stable Unit Treatment Value Assumption (SUTVA)).[28,29] Second, we assume that conditional on the potential confounding variables W, treatment assignment, and perception are independent of the outcome ($A, P \perp Y_{a,p}|W, \ \forall a, p$). This assumption is referred to as the "randomization assumption" or the "no unmeasured confounding assumption" so that W is assumed to contain all variables that cause both P and Y. Finally, we assume that each treatment/perception combination (a, p) is possible for all the members of the target population, i.e. $P(A = a, P = p|W) > 0 \ \forall W$. This last condition is referred to as the "experimental treatment assignment" assumption or the positivity assumption.[30,31]Given the graph in figure 2.1, the likelihood of an observed data observation can be factorized as: $l(O) = P(W)P(A|W)P(P|A, W)P(Y|A, W, P) = P(W)g(A, P, W)Q(Y, A, P, W)$ where $g(A, P, W) \equiv P(A|W)P(P|A, W)$ and $Q(A, P, W) \equiv E[Y|A, W, P]$.

2.7 Maximum Likelihood Estimation

The MLE works specifically with the term $Q(A, P, W)$, that may be estimated using an appropriate regression model; this approach does not require estimates of distributions defined by the terms that determine treatment and perception distribution. Counterfactual distributions of the data under specific interventions are defined by the G-computation formula. The name, G-computation, stands for graphical computation and has roots in graphical modeling. Assuming a particular causal graph, the likelihood can be factorized, and the G-computation formula is obtained by carrying out a specific intervention on the likelihood. The obtained

formula represents the counterfactual distribution the data would have had under the specific intervention. The MLE is sometimes referred to as the G-computation estimator.

$$\Pr(Y_a < y) = \int \Pr(Y < y | A = a, P = p, W)\, dF(w)$$

Using the G-computation formula, estimates of the relevant population counterfactual means are $E\left[Y_{a,p}\right] = E_W[E[Y|A = a, P = p, W]]$, and thus our counterfactual means of interest can be estimated by $\hat{E}\left[Y_{a,p}\right] = \frac{1}{n}\sum_{i=1}^{n} \hat{E}[Y|A = a, P = p, W_i]$, where i indexes participants, $i = 1, \ldots n$. Each comparison, defined in Sections 4 and 5, may be estimated by first estimating each of two defining marginal expectations and then taking their difference. For instance, for a dichotomous treatment variable, A, $\psi_a(p) = E[Y_{1,p}] - E[Y_{0,p}]$ may be estimated by

$$\begin{aligned}\hat{\psi}_a(p) &= \hat{E}[Y_{1,p}] - \hat{E}\left[Y_{0,p}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \hat{Q}(A = 1, P = p, W_i) - \frac{1}{n}\sum_{i=1}^{n} \hat{Q}(A = 0, P = p, W_i)\end{aligned} \quad (1)$$

The MLE estimator $\hat{Q}(A, P, W) \equiv \hat{E}[Y|A, W, P]$ can be based on a parametric model; however the consistency of this estimator relies on the consistency of the regression. Given that nonparametric approaches are not feasible if W is high-dimensional, we suggest use of a machine-learning algorithm that allows the user to specify the degree of flexibility in regression terms and includes some form of model selection. In particular, in the example of Section 7 we employ the Deletion Substitution Algorithm (DSA)[32] to choose the final regression form. DSA is a data-adaptive model selection algorithm based on cross-validation. The algorithm selects from a set of candidate generalized linear models that consist of polynomials of the covariates and their tensor products. The candidate models are produced by three different moves: deletions, substitutions, and additions. A deletion step removes a term from the model, a substitution step replaces a variable with another, and an addition step adds a variable to the model. The final model selected by the DSA minimizes the empirical risk on the learning set. [32] The algorithm limits the search for the best model through user specified parameters for the space of candidate models such as the maximum sum of powers for the variables and the maximum order of interaction between them. Standard errors (and confidence intervals) that account for data-driven levels of flexibility in the regression model can be based on the bootstrap. The hope for using this model and a simple plug-in estimator for our parameter of interest is that the optimization in the balance between bias and variance in the estimate of the regression model using this approach will translate to a close to optimal variance-bias trade-off for the parameter estimate of interest.[33] However, in section 8 we will discuss a generalization of this approach that more directly targets this model selection towards estimating the parameter of interest.

For the gabapentin trial, parameters of sections 4 and 5 were estimated using Maximum Likelihood Estimation of the G-computation formula, which requires modeling of the end-point pain score on treatment, perception and baseline covariates, $E[Y|A,P,W]$. The model was selected using the machine-learning, DSA algorithm with perception, treatment, and the interaction between the two, forced into the model, with the rest of the terms selected from basis functions of the following measured confounding variables: sex, age, race, height, weight, baseline pain score, and baseline sleep score. The covariates selected in the model were restricted to no higher than second-degree terms and, similarly, only allowed for two-way interactions. Five-fold cross-validation was used within the DSA algorithm for model comparison and selection.

Once the regression model has been fit, the marginal mean $E[Y_{a,p}]$ is estimated by using the final fitted regression model to predict, for each individual, their outcome keeping the covariates as observed, but fixing treatment at $A = a$ and perception at $P = p$; this yields, for the i^{th} observation, the "predicted" value $\hat{E}[Y|A = a, P = p, W_i]$. These predicted outcomes were then averaged over individuals (and thus over the empirical distribution of W) to give $\hat{E}[Y_{a,p}] \equiv \hat{E}_W[\hat{E}[Y|A = a, P = p, W]$ as discussed in Section 6. Finally, to estimate the marginal effect $E[Y_{a,p}] - E[Y_{a',p'}]$, we simply subtracted $\hat{E}[Y_{a',p'}]$ from $\hat{E}[Y_{a,p}]$. Bootstrap standard errors were estimated by re-sampling the observations with replacement 5000 times, performing model selection using DSA for every bootstrap sample, and finally, estimating the desired parameters as above. Visual checks on the bootstrap distributions showed symmetric distributions of the bootstrap estimates around the full-data estimate, suggesting that the variability introduced by the model selection is of 2nd order; note that examination of the bootstrap distribution provides an informal diagnostic on whether the estimate has the desired sampling distribution. For each parameter, the estimated standard error was used to calculate a two-sided Wald test statistic and a subsequent p-value. For the original data, the following variables were selected by DSA for the estimator (in addition to perception, treatment, and their interaction, that were forced into the model): baseline pain score, baseline sleep score, and the second power of baseline sleep score. We note that the chosen model for any flexible machine learning algorithm with a relatively small sample size is unstable in its 'choice' of included covariates, and so one can conclude very little about the relative importance of variables from a single fit.

Table 2.1 shows the resulting MLE estimates for four comparisons: the treatment effects at both levels of perception, and the perception effects at both levels of treatment. For example, based on these results, the average effect of gabapentin on endpoint pain scores, with P fixed at no knowledge of treatment or placebo (i.e. in this case, no side effects), is estimated to be $\hat{\psi}_a(0) = 0.71$ points, with an associated 95% confidence interval of (-0.14, 1.56). The estimated effect of gabapentin with perception set to $P = 1$ (i.e. everyone had side effects) is $\hat{\psi}_a(1) = 2.50$ points, with a 95% confidence interval of (1.15, 3.84).

Table 2.1 Estimated parameters using MLE and the corresponding 95% confidence intervals

Parameter	G-comp Estimate	P-value	95 % CI
$\psi_a(0) = E[Y_{0,0}] - E[Y_{1,0}]$	0.71 (0.44)	0.10	(-0.14,1.56)
$\psi_a(1) = E[Y_{0,1}] - E[Y_{1,1}]$	2.50 (0.69)	0.0002	(1.15,3.84)
$\psi_p(0) = E[Y_{0,0}] - E[Y_{0,1}]$	-0.59 (0.64)	0.35	(-1.83,0.65)
$\psi_p(1) = E[Y_{1,0}] - E[Y_{1,1}]$	1.18 (0.0.51)	0.02	(0.18,2.17)

The average perception effect with treatment fixed at $A = 0$ (placebo) is $\hat{\psi}_p(0) = -0.59$ with a 95% confidence interval of (-1.83, 0.65), and, conversely, the estimate of the perception effect with treatment fixed at $A = 1$ (gabapentin) is $\hat{\psi}_p(1) = 1.18$ with a 95% confidence interval of (0.18, 2.17). That is, the results imply that, contrary to fact, had everyone received the treatment, patients with treatment-related side effects are estimated to report significantly greater pain reduction than if they had no side effects. This in turn suggests that treated patients who believe that they are receiving active treatment report significantly greater pain reduction, on average, than treated individuals that have no opinion about their treatment.

In summary, we estimate about 40% of the naïve estimated treatment effect (ignoring treatment-related side effects) disappears if no one would have experienced side effects and presumably remained unbiased in reporting their pain scores. That is, after accounting for perception, the estimated mean differences suggest that gabapentin does not have a statistically significant effect on pain reduction had no individual experienced a treatment-related side effect during the trial: the estimated treatment effect with P fixed at 0 is no longer significant. However, the treatment effect amongst those with side effects is much higher and highly statistically significant. There are several possible explanations for this result that are all consistent with the data. The obvious explanation is that unmasking creates the very bias that blinding is designed to protect against; an alternative hypothesis is that the occurrence of treatment-related side effects is an indication or proxy that the drug is having an efficacious effect and it is exactly this group of individuals who experience pain reduction. Unfortunately, the data cannot possibly distinguish between these two alternative interpretations. However, given that masking in subjectively scored pain trials is considered so key to obtaining unbiased results, it would be inappropriate in our view to assign the entire treatment effect when $P = 1$ to a therapeutic effect of the drug.

Some participants were lost to follow up before the end of the study (14 and 18 patients in the gabapentin and placebo groups, respectively). We followed the investigators in using the average of the last seven pain measurements as indicated, the so-called "carry-forward" method. In both treatment and placebo groups, there is a tendency for the pain scores to decline over the eight week follow-up period. This means that the 'carry-forward' approach to missing data for those lost to follow-up tends to lead to overestimates of the mean endpoint pain scores in both treatment and

placebo groups. To investigate the sensitivity of the findings to this form of imputation, we did an equivalent analysis using only participants who remained under follow-up the entire eight weeks. Effectively this assumes that the data for individuals lost to follow-up is missing at random. Although the standard errors are necessarily higher, there is a marked difference in the estimated average treatment effects at both levels of P. Specifically, we now estimate $\hat{\psi}_a(0) = -0.05$ points, and $\hat{\psi}_a(1) = 1.53$. Thus, the entire treatment effect is now erased when unmasking (i.e. occurrence of side effects) is accounted for; in fact the treated patients are very slightly worse off than those on placebo when we set $P = 0$. The estimated treatment effect amongst the unmasked participants ($P = 1$) is still notable but smaller than when the carry-forward method is employed, $\hat{\psi}_a(1) = 2.50$.

2.8 Targeted Maximum Likelihood Estimation

We discussed a MLE estimator in the previous section, but there also exist estimating equation approaches for these types of parameters in semi-parametric models. Specifically, the Inverse Probability of Treatment Weighted (IPTW) estimator[20,21] represents a different approach to estimation of the causal parameters of interest here; this approach requires estimation of the treatment/perception assignment mechanism as determined by $g(A, P, W)$. In addition, this estimator can be augmented such that the new estimator is double robust (so-called Double Robust Inverse Probability of Treatment Weighted, or DR-IPTW estimator[22,23,24]). The virtue of this estimator is that is consistent if either the outcome regression model or the treatment/assignment mechanism is correctly specified. In addition, the estimator is locally efficient, so it achieves (under assumptions) maximal efficiency among competing estimators in a semi-parametric model. However, often these estimating equation approaches have very poor finite sample performance, which for instance may result in the estimator not necessarily being bounded between the natural limits of the parameter of interest (e.g., probability differences < -1 or > 1). Optimally, one would like the asymptotic properties of the DR-IPTW estimator, but with the finite sample virtues of the MLE estimator. This is achieved by targeted maximum likelihood estimation.

What machine learning approach alone lacks is that the algorithm is not optimized towards the parameter of interest; whereas it might provide an optimal estimator of the prediction of Y, given A,P,W, it may be a poor estimator for a particular parameter that is a function of this model.[25] Typically, a plug-in estimator for the parameter of the density estimator will be biased due to model misspecification (unless the estimate is nonparametric) as noted above for the MLE. The standard criteria for model improvement (by making it more flexible) usually focus on the model and not the ultimate parameter of interest. In such cases, the TMLE directly addresses the bias issue by carrying out a subsequent clever parametric maximum likelihood fit that is directly tailored to remove bias for the target parameter of interest, treating the initial MLE estimator as an offset. In particular, the TMLE modifies maximum likelihood estimation in a way that yields a plug-in estimator with the influence curve equal to the efficient influence curve.[25] Beyond its asymptotic

efficiency, the resulting estimator will also be double robust (in fact it will be so-called collaboratively double robust – see Van der Laan and Gruber[33]). Practical consequences could be that covariates that are potential confounders (W) for the association of interest might be dropped by a model selection procedure for estimating Q as we are simply trying to get the best density estimate of Y given A, W, P (rather than estimation of regression coefficients). To retrieve robustness, TMLE works by augmenting the original fit of this density (i.e., the conditional distribution of Y, given A, W) by adding an appropriate 'clever' covariate (the choice of this covariate, relative to the choice of the models for Q and g is discussed in Van der Laan and Rubin[25]).

As noted, the estimator is consistent if either $Q(A, P, W)$ or $g(A, P, W)$ is consistently estimated. A model selection algorithm such as the DSA[32] may be used for finding the initial estimate Q_n^0 as before. Given this initial estimate, the updated Q is given by (in the linear model):

$$Q_n^*(A, P, W)) = Q_n^0(A, P, W) + \varepsilon h(A, P, W) \quad (2)$$

where the derived covariate $h(A, P, W)$ is a function of $g(A, P, W)$.[25] Here ε is estimated using maximum likelihood (e.g. least squares), where Q_n^0 is treated as an offset, as further discussed below. For the defined mean parameters $E[Y_{a,p}]$ of Sections 4 and 5, $h(A, P, W)$ is equal to $\frac{I(A=a,P=p)}{P(A=a|W)P(P=p|A=a,W)}$. If one wishes to solely target a difference such as $E[Y_{1,p} - Y_{0,p}]$, one can use as single clever covariate the difference of the two corresponding clever covariates. To implement TMLE, we have to estimate the denominator of $h(A, P, W)$. For the first term, $P(A = a|W)$, we use the fact that A is randomized (independent of W) and substitute either the known treatment assignment probabilities, or, to gain efficiency[25], the estimated empirical treatment assignment proportions ignoring W. The second term requires estimation through some form of binary (such as logistic) regression as perception presumably varies by treatment group and possibly the added covariates W; model selection can be employed here to allow as much flexibility in estimation as supported by the data. In observational studies for which treatment A is not randomized, the treatment mechanism may also be modeled using a logistic regression similar to the model for perception.

Once the h term is approximated, the coefficient ε is estimated using maximum likelihood for the regression model (2) assuming Gaussian errors; here Q_n^0 is a fixed offset in the model and ε is the coefficient to be estimated. The magnitude of the estimated ε depends on the amount of residual confounding (for estimation of the targeted mean parameter) along the direction of $h(A, P, W)$. This process is iterated until convergence. In our simple case, convergence is achieved in one step: see Van der Laan and Rubin[25].

This estimator is identical to estimator given in (1) except that it uses the updated Q^*, instead of Q, with the consequence that bias is reduced by directly targeting the desired mean and double robustness is attained. Estimation of each mean requires separately updated $\hat{Q}^*$s, for example, one for $\hat{Q}(A = 1, P = p, W)$, and a second for $\hat{Q}(A = 0, P = p, W)$. Each of the updated $\hat{Q}^*$s is obtained by plugging in the corresponding choice of $A = a, P = p$ to determine the appropriate 'clever covariate'. Under regularity conditions, TMLE is a consistent and asymptotic linear estimator.[18] Consequently, for inference and testing we may use $\hat{\psi}_a \sim N(\psi_a, \Sigma_0)$, where ψ_a is the true population parameter, and Σ_0 is the variance of the efficient influence curve. The latter variance can be estimated via the sample variance of a plug-in estimate of the efficient influence curve, or more safely by using the bootstrap as illustrated in the following section.

Any combination of the mean parameters can then be directly estimated from the TMLEs of the individual components. Of most interest, for example, the average causal effect of treatment had everyone's perception remained at a fixed level $P = p$ can be estimated by:

$$\hat{\psi}_a(p) = \hat{E}[Y_{1,p}] - \hat{E}[Y_{0,p}] = \frac{1}{n}\sum_{i=1}^{n} \hat{Q}^*(A = 1, P = p, W_i) - \frac{1}{n}\sum_{i=1}^{n} \hat{Q}^*(A = 0, P = p, W_i) \quad (3)$$

A TMLE for a specific mean comparison can be computed, specifically targeting this parameter of interest. We do not pursue this further here as we wish to examine several mean comparisons simultaneously.

Following our earlier analysis of the gabapentin trial, we used TMLE as an augmentation to the MLE with the possibility of minimizing bias due to model misspecification, as well as reducing the variability of estimation. As the first step in estimating each of the four means, $E[Y_{a,p}]$, we used the MLE estimator of Section 7.3. To approximate the 'clever' covariate, the empirical proportion of treated patients was used instead of using 0.5 for probability of treatment $P_0(A = a)$. The probability that $P = 1$ was modeled based on treatment and the baseline covariates as discussed in Section 6.1: the DSA algorithm was used to fit a logistic regression model with treatment forced into the model. Once again, predictors were restricted to second degree, and interaction terms restricted to two-way terms, with five-fold cross validation used for model comparison and selection. Subsequently, we used precisely the same plug-in estimator, with the model of $E[Y|A, P, W]$ augmented by the clever covariate noted in Section 6.1. As discussed above, we "targeted" each of the estimated marginal means $\hat{E}[Y_{0,0}]$, $\hat{E}[Y_{0,1}]$, $\hat{E}[Y_{1,0}]$, and $\hat{E}[Y_{1,0}]$ separately. As above, the marginal difference $E[Y_{a,p}] - E[Y_{a',p'}]$, was simply estimated by subtracting $\hat{E}[Y_{a',p'}]$ from $\hat{E}[Y_{a,p}]$. Bootstrap standard errors were calculated by repeating the TMLE fit to each of 5000 bootstrap samples. For each bootstrap sample, the clever covariate was formed separately based on the treatment mechanism model (g) selected for that particular sample. Regression models for each sample (Q) were

augmented by the corresponding clever covariates, the responses were predicted for each individual, and the marginal means estimated as before. Once again, a two sided Wald test was used to calculate p-values.

Table 2.2 provides the analogous estimates to Table 2.1, but now using TMLE. The estimates are generally similar to the un-augmented results, reflecting little residual bias with respect to the observed covariates in the latter method. There are increases in the standard errors associated with TMLE results, possibly due to finite sample issues with regards to the method used to estimate g. With regard to the latter point, for a few levels of W, the model predicts small probabilities for perception (= *1*), given treatment, thereby inflating the variance of the 'clever' covariate and, subsequently, the predicted pain scores. Though beyond the scope of this paper, one can construct the 'clever' covariate in a manner that can ameliorate this problem.[21] In summary, the approach here confirms the results and interpretation achieved through MLE approach.

An additional efficiency enhancement could result from methods used to properly constrain models to predict the outcomes within a known limited range (as here where the pain score must lie between 0 and 10). Specifically, one can transform the dependent variable to lie between 0 and 1, and use a TMLE logistic regression approach that guarantees that all predicted scores fall within the known range[34]. This enhancement will be demonstrated in the next chapter for the longitudinal setting. One thus has several tools at their disposal that can help to optimize finite sample performance.

Table 2.2 Estimated parameters using TMLE and the corresponding 95% confidence intervals

Parameter	TMLE Estimate	P-value	95 % CI
$\psi_a(0) = E[Y_{0,0}] - E[Y_{1,0}]$	0.78 (0.58)	0.18	(-0.35,1.91)
$\psi_a(1) = E[Y_{0,1}] - E[Y_{1,1}]$	1.98 (0.99)	0.04	(0.04,3.91)
$\psi_p(0) = E[Y_{0,0}] - E[Y_{0,1}]$	-0.07 (0.81)	0.93	(-1.64,1.50)
$\psi_p(1) = E[Y_{1,0}] - E[Y_{1,1}]$	1.12 (0.62)	0.07	(-0.09,2.32)

2.9 Discussion

It is important to note an asymmetry in our analysis regarding the use of observable treatment-related side effects as a proxy for treatment perception or unmasking. That is, we have not accounted for the possibility that some participant's perception changes to $P = -1$ (convinced they are on placebo) during the course of the trial. This might occur because an individual feels no benefit from treatment thereby becoming convinced that they are on placebo, and possibly biasing their subjective endpoint pain scores upwards. That is, there is no analogue of the potential unmasking side effect in the other direction. It is possible that adjusting for this additional unmasking effect might increase treatment efficacy estimates were

unmasking maintained ($P = 0$) although this is impossible to determine in the absence of further information.

This observation, and the fact that we cannot disentangle a significant perception effect from a stronger treatment effect for those who had the relevant side effects, suggest the need for collecting data on patients' perception in randomized trials with subjectively recorded outcomes, previously suggested by other authors including Turner et al.[35] We note that in some vaccine efficacy trials, investigators have been concerned about patients increased risk behavior due to their treatment perception and have suggested collecting data on perception. In a hepatitis B vaccine efficacy trial, placebo recipients were at a higher risk of hepatitis B infection after their final injection. This higher risk of infection suggested that some placebo recipients may have assumed they were protected and increased their risk behavior.[36] Similarly, Bartholow et al.[37] investigated sexual risk behavior of participants in an HIV vaccine efficacy trial, and found that among younger (<30) men who have sex with men, perceived assignment to vaccine was associated with an increased probability of unprotected sex.

As we demonstrated in this chapter, investigators may estimate and remove some adverse post-randomization confounding factors of the effect of treatment using the causal inference framework. Suitable definition of appropriate parameters and the use of semi-parametric machine-learning techniques allow investigators to obtain less biased and more interpretable estimates. In particular, TMLE estimation of parameters defined by the G-computation formula improves upon the MLE approach by reducing bias and improving efficiency of estimators, with additional enhancements available to improve finite sample performance. In the example, the methods demonstrate that the naïve approaches to accommodate treatment-related side effects and using carry-forward to impute missing data both considerably distort the assessment of treatment efficacy.

In the gabapentin example we focused on placebo, perception, and unmasking effects. However, any other measurable non-specific effect of the treatment may also be estimated using the same framework. Perception effects are non-specific effects of treatment and therefore a component of the placebo effect. As an alternative to adjusting for perception effects in subjective outcome clinical trials, we suggest conducting (i) randomized trials with no-treatment control groups to estimate the placebo effect as a whole (in possible settings), and/or (ii) small unmasked randomized trials parallel to the main trial to allow direct estimation of the masking effect.

We note that the desired parameters of interests, and appropriate estimation techniques, become more complex when components of W are on the direct causal pathway describing the effect of treatment on the outcome. This is not the case for any covariates in the gabapentin example discussed in Section 7. Targeted maximum likelihood has been recently extended to accommodate this situation.[38,39]

The methods and analysis proposed in this chapter did not exploit the particular time when side effects and possible unmasking occurs during the conduct of the trial. Longitudinal observations provide such additional information and can be used to refine the techniques outlined here. In the next chapter we will demonstrate how the suggested methods can be extended to the analysis of longitudinal information on unmasking.

Chapter 3

Perception and Unmasking Effects in Longitudinal Settings

3.1 Background

In the second chapter, we used a counterfactual framework from causal inference literature to formally define a placebo effect and some of its components such as unmasking and perception effects in a cross-sectional setting. We defined direct effects of a treatment under hypothetical interventions on patients' masking and perception; had the patients remained masked throughout the trial, or had the patients' perception regarding treatment been kept at a fixed level. This framework was applied to a pain trial in which occurrence of treatment-related side effects were used as a proxy for unmasking of the patients. Semi-parametric estimation methods such as Maximum Likelihood Estimation and Targeted Maximum Likelihood Estimation were utilized for obtaining estimates of treatment and perception effects. In this chapter, we generalize the concepts of perception and unmasking effects to longitudinal clinical trials. We reintroduce the gabapentin trial as a longitudinal study and explain the causal framework and the generalization of the assumptions required to estimate causal effects for the longitudinal setting. The parameters of interest will be defined as in the previous chapter and the generalization of the two estimation methods (MLE and TMLE) will be used to estimate these parameters.

3.2 Time Dependent Perception and Unmasking

In a longitudinal setting, patients' perception regarding their treatment may vary as a function of time. For instance, a patient may believe he is on a placebo during the early stages of the trial due to a slow effect of the treatment but may change his perception later on as a result of observing treatment related side effects. We refer to this time dependent random variable of degree of certainty as a patient's *perception at time* t, $P(T = t)$, where $P(t) = 1$ indicates that a patient is certain that he has been assigned the active treatment at time t, and at the other end of the scale, $P(t) = 0$ indicates that a patient is certain that he is receiving the placebo at time t.

At any time point during the course of the trial, the investigator may inform the patients of their treatment groups. In this situation, the patients would automatically have their perception $P(t) = 1$, or $P(t) = 0$, for the rest of the trial ($T \geq t$). If the investigator informs the patients of their true treatment group, the patients are considered to be unmasked. However, having perception $P(t) = 1$, or $P(t) = 0$ does not directly imply unmasking of a patient since the patient may be incorrectly convinced of his treatment assignment. On the other hand, unmasking of a patient

implies that his perception $P(t) = 1$, or $P(t) = 0$, depending on the treatment assignment. In other words, unmasking may be defined as a patient having $P(t) = 1$, and having treatment assignment $Z = 1$, or having $P(t) = 0$ and treatment assignment $Z = 0$.

As in the cross-sectional setting, we focus on a discrete approximation for a patient's perception. In particular, we consider a three-level variable,$P(t)$, for perception level at time t. $P(t) = 1$ would indicate that a patient believes he is on treatment at time t, $P(t) = 0$ would denote a patient having no knowledge of the treatment, and finally, $P(t) = -1$ would show that a patient believes he is on a placebo.

3.3 A Causal Inference Framework for Longitudinal Settings

To formulate causal effects of interest for time dependent settings, we follow the counterfactual framework used in the second chapter. This framework was first considered by Neyman[40], and later revisited by Rubin[41], Robins [42], and Holland[43]. Direct effects have been defined by Robins and Greenland[44], Pearl[17], and Robins[45] under a general framework which regards observed data as a missing data structure from a full data structure consisting of all the potential counterfactuals for the intermediate variables and the outcome for every possible treatment. Under this framework, the full data structure for the gabapentin trial would consist of all pain measurements for different hypothetical combinations of treatment and perception for each patient through time. (also see Peterson et al.[46], and Rosenblum et al.[19]). Any causal effect of interest may then be defined as a difference in the counterfactual outcomes under two hypothetical treatments. To identify such causal effects, it is generally assumed that such counterfactuals exist. For instance, one can imagine a hypothetical experiment in which the gabapentin trial investigators had randomized the patients to both treatment and perception groups. Even though such hypothetical interventions may not be ethical in some cases, we will consider them here for demonstrational purposes.

Consider the gabapentin trial in which baseline covariates are measured and the patients are randomized to two different treatment arms, placebo$(A = 0)$, and active treatment$(A = 1)$. The investigator asks the patients to rate their pain every day on a scale of 0 to 10, and monitors the patients' side effects for the duration of the trial. At the end of the trial the investigator has observed the following chronological data structure (assuming full compliance to the assigned treatment):

$$O = (W(0), L(0), A(0), P(1), L(1), P(2), L(2), \dots, P(K), L(K), Y = L(K+1))$$
$$\text{for } T = 0, \dots, K = 1$$

where $L(t)$ is the patient's pain measured at time t, and $P(t)$ is patient's time dependent perception regarding his treatment, measured as binary variable (treatment/ placebo or no perception).

The gabapentin trial design could have been modified to include an intervention on perception or unmasking of the patients. As a simple example, assigning a binary treatment at the start of the trial and a binary perception at a particular time $T = t$, would have resulted in four joint treatment and perception groups. Within those assigned to treatment, $A = 1$, the subgroup assigned to perception $P(t) = 1$, would have been told that they were receiving the active treatment, at time $T = t$, and the subgroup randomized to perception $P(t) = 0$, would have been told that they were receiving a placebo, at time $T = t$. Similarly, for the patients assigned to placebo, $A = 0$, the subgroup assigned to perception $P(t) = 1$, would have been told that they were receiving the active treatment, at time $T = t$, and the subgroup randomized to perception $P(t) = 0$, would have been told that they were receiving a placebo, at time $T = t$. If the information given to patients was regardless of their true treatment assignment the investigator would have observed the patients' counterfactual outcome for the assigned perception. On the other hand, if the information given to the patients was based on their true treatment assignment, the investigator would have observed the patient's counterfactual outcome for the assigned masking status.

In a longitudinal setting, the full and the observed data structure are defined in the following manner; for convenience in notation, let $\bar{Y}(t) = (Y(s): s \leq t)$ be the time dependent history of a variable Y up to, and including time t. To define the full data structure, let $\mathcal{P}$ denote the set of all possible perception histories, and let $\mathcal{A}$ be the set of all possible treatments. For every $a \in \mathcal{A}$, and $\bar{p} \in \mathcal{P}$, let $L_{a,\bar{p}} \equiv (L_{a,\bar{p}}(t): t = 0, \ldots, K+1)$ denote the treatment specific outcome process one would have observed if the patient would have followed treatment $A = a$, and his perception process would have been controlled at $\bar{P} = \bar{p}$. Then $X_{a,\bar{p}}(t) \equiv (A = a, \bar{P}_a(t), L_{a,\bar{p}}(t): a, \bar{p})$ is the complete collection of counterfactual data structure for treatment and perception specific history we would have observed if the patient would have received treatment $A = a$, and his perception controlled at $\bar{P} = \bar{p}$. To define the observed data structure as a subset of the full data structure, let $A = A(0)$ be the treatment assignment at time $t = 0$, and let $\bar{P} = \bar{P}(t) = (P(0), \ldots, P(K))$ be the observed perception history up to time and including time K. In addition, let W be the set of baseline variables, and let $\bar{L}(t) = (L(0), \ldots, L(K+1))$ be the observed pain history up to and including time $K + 1$. Then the observed data may be rewritten as $O = \Big(W, A, \bar{P} = \bar{P}_A(K), \bar{L} =$ *LA,PK+1.*

The full data and the observed data structures are tied together by assuming that the observed pain, $L_{A,\bar{P}}(t)$, is equal to the treatment-specific pain, $L_{a,\bar{p}}(t)$, corresponding with the treatment and perception process the subject actually followed, $L_{a,\bar{p}}(t) = L_{A,\bar{P}}(t),\ t = 0, \ldots, K+1$. This assumption is known as the "consistency assumption" or the Stable Unit Treatment Value Assumption (SUTVA).[28,29,38] For this assumption to hold, the observed outcome of a patient cannot be affected by the outcome of any other patient. In the context of the gabapentin trial, the consistency assumption states that the pain outcome observed for a patient who had received treatment and was

unmasked at time t (perhaps due to observed side effects), would have been equal to the hypothetical outcome had the investigators told the patient that he was receiving treatment at time t.

3.4 Longitudinal Direct Effects and the Parameters of Interest

Under the framework of the previous section, Robins and Greenland[44], Robins[45], and Pearl[17] define the direct effect of the treatment in the following manner:

$$\psi_a(\bar{P}(t) = \bar{P}_0) = E[L_{a,\bar{P}_0} - L_{0,\bar{P}_0}] = E\left[L_{a,\bar{P}_0}\right] - E[L_{0,\bar{P}_0}]$$

Where, $\bar{P}_0$ is the perception counterfactual had the patient received a placebo, $L_{a,\bar{P}_a} - L_{0,\bar{P}_0}$ is an individual direct effect of the treatment, and the population direct effect of treatment is defined as the mean of the individual direct effects. For the gabapentin trial, the type I direct effect of the treatment is defined as the difference in the pain outcome had all the patients received treatment and their perception set at its value under no treatment, and the outcome had all the patients received no treatment, and their perception fixed at its value under no treatment.

A direct effect of the treatment for perception fixed at $\bar{P}_0$ may not always be the treatment parameter of interest. In a more general sense, the investigator may be interested in treatment effects under any specific perception patterns $\bar{p}$.

$$\psi_a(\bar{P}(t) = \bar{p}) = E[L_{a,\bar{p}} - L_{0,\bar{p}}] = E\left[L_{a,\bar{p}}\right] - E[L_{0,\bar{p}}]$$

The above parameterization allows the investigator to answer questions such as the effect of the treatment had all the patients been kept masked throughout the trial or had they been unmasked halfway through the course of the trial.

In addition to a treatment effect for a particular perception pattern, the investigator may be interested in the effect of a perception pattern for a fixed treatment level. In general, we define longitudinal perception effects based on two distinct perception patterns $\bar{p}_i, \bar{p}_j$ for $\neq j$. The perception effect on the outcome, due to having perception pattern $\bar{P} = \bar{p}_i$ compared to $\bar{P} = \bar{p}_j$, at fixed treatment level a, may be defined as $\psi_p(\bar{P}_1 = \bar{p}_i, \bar{P}_2 = \bar{p}_j, A = a) = E[Y_{a,\bar{p}_i}] - E[Y_{a,\bar{p}_j}]$;

For such parameters to be consistently estimated we need an additional set of assumptions. It must be assumed that there are no unmeasured confounders of the effect of treatment on pain (treatment is randomized) and no unmeasured confounders of the effect of perception at every time point t on pain given the past. In the gabapentin trial where treatment is randomized, this assumption is referred to as the "randomization assumption".[28] For the case of perception at each time point t which is not randomized, the assumption is referred to as "no unmeasured confounding".[28]

$$A \perp Y_{a,p}, \text{ and } P(t) \perp \{Y_{a,p} | W, A, \bar{P}_{A=a}(t-1), \bar{L}_{A,\bar{P}}(t-1)\}$$

An additional assumption is the Experimental Treatment Assumption (ETA) for treatment and perception. For treatment, $\mathcal{P}\big(A(0) = 1 \big| W(0)\big)$does not equal to 0 or 1,

and for perception at time t, $\mathcal{P}(P(t) = 1|W, A, \bar{P}_{A=a}(t-1), \bar{L}_{A,\bar{P}}(t-1))$ does not equal to 0 or 1, almost surely, for any values of $W, A, \bar{P}_{A=a}(t-1), L_{A,\bar{P}}(t-1)$.[28] Finally, it is assumed that a patient's pain process is not affected by his perception level after the pain process is measured, $L_{a,\bar{p}}(t)$=$L_{a,\bar{p}(t-1)}(t)$. This assumption is referred to as the "time ordering assumption".[28]

3.5 Causal Graphs and Additional Model Assumptions

In this section, a few plausible longitudinal causal graphs are compared for the gabapentin trial. For simplicity, we restrict our discussion of the causal graphs to three time points only ($t = 0,1,2$). Later, we will demonstrate that considering additional time points for analyzing the gabapentin data requires more observations and we are unable to draw reasonable conclusions due to the curse of dimensionality. Baseline pain,$L(0)$, and other baseline covariates are measured and the patients are randomized to either treatment or placebo ($A(0) = 0,1$) during week 0 of the trial, at time $t = 0$. The end of the titration period at the end of the 4th week corresponds to time $t = 1$, and time $t = 2$ represents the conclusion of the trial at the end of the 8th week. Thus, $L(1)$ and $L(2)$ represent average pain measurements during weeks 4 and 8 respectively. The perception of the patients for the first half of the trial will be denoted by $P(1)$ and is defined as an indicator for presence of any treatment related side effects up to the end of week 3. Similarly, the perception of the patients at the end of the trial is denoted by $P(2)$ and is defined as the presence of any side effect up to the end of week 7. Based on our definition of perception and pain variables, we observe the following simplified chronological order:

$$L(0), A(0), P(1), L(1), P(2), L(2)$$

Treating the gabapentin trial as a cross-sectional study and using presence of any side effects during the course of the trial as a proxy for perception P ignores both the time of occurrence of side effects and the intermediate response measurements. The equivalent longitudinal model (figure 3.1) with three time points only incorporates the last observed perception $P(2)$. In this model, baseline covariates can affect both a patient's final perception and the outcome. Yet, the model allows formulation of a direct effect of treatment assignment on the pain outcome, and an indirect effect of the treatment through perception.

A simple generalization of this cross-sectional model would be to incorporate the effect of perception at time $t = 1$ in addition to the final perception level at time $t = 2$. For this improved model (figure 3.2), treatment and earlier perception affect the outcome both directly and indirectly. The indirect effect of treatment may be mediated through $P(1)$, $P(2)$, or through a combination of $P(1)$ and $P(2)$. Similarly, $P(1)$ may either affect the outcome directly, or have an indirect effect on $L(2)$ through $P(2)$.

Even though the improved cross-sectional model incorporates perception for all time points, it assumes that $P(1)$ and $P(2)$ affect the pain response directly and do not affect intermediate pain measurements. Relaxing this assumption and allowing the perception effect to be mediated through the intermediate pain variable, $L(1)$, we obtain the causal graph of figure 3.3.

Figure 3.1 A cross-sectional causal graph for the gabapentin trial

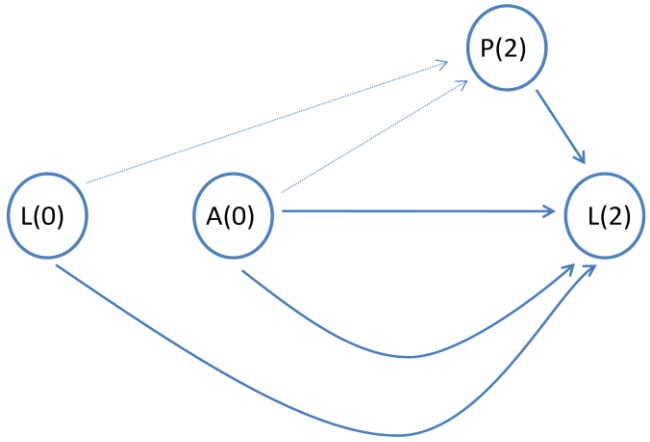


Figure 3.2 A causal graph for the gabapentin trial incorporating time of occurrence of side effects

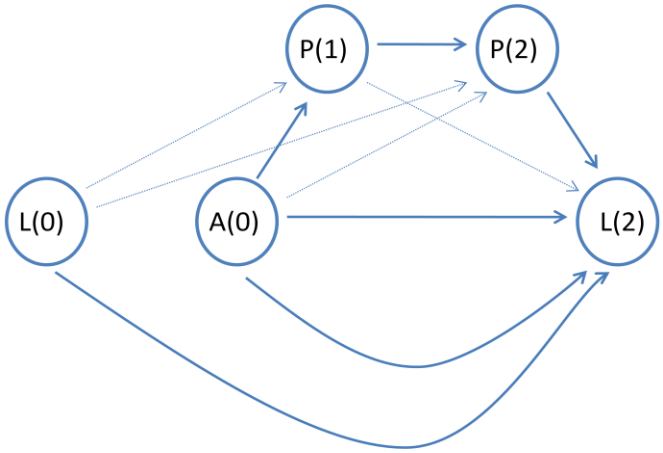
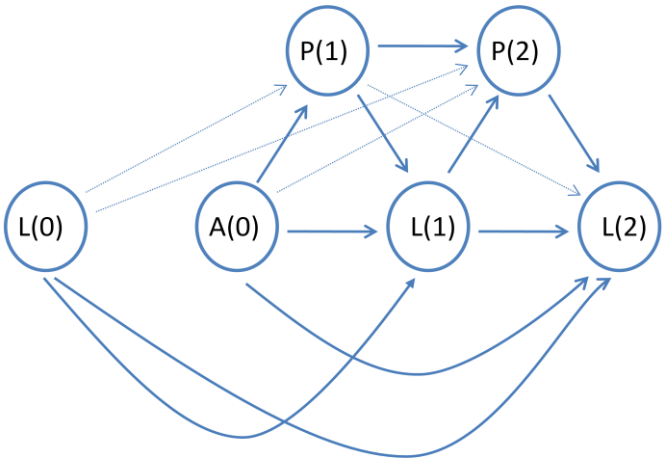


Figure 3.3 A causal graph for the gabapentin trial incorporating time of occurrence of side effects and intermediate pain score



In this graph (figure 3.3), intermediate pain measurement $L(1)$ confounds the effect of $P(2)$ on final pain and should be adjusted for if we are interested in the effect of $P(2)$ on pain. However $L(1)$ is affected by both treatment $A(0)$ and perception $P(1)$ and adjusting for it will block the causal pathways from these two nodes to the outcome. Fixing $L(1)$ is equivalent to assuming that perception $P(1)$ does not affect intermediate pain $L(1)$, that pain $L(1)$ does not affect $P(2)$, and that there is no indirect effect of treatment through $L(1)$. Hence, classical longitudinal models fail to capture the relations portrayed in this final model, and the investigator needs to rely on other methods which will allow estimation of the direct effects of treatment and perception on the pain outcome.

For the remainder of the chapter, we will focus on estimating direct and indirect effects of treatment and perception based on the above longitudinal model. This model accounts for the time of occurrence of side effects and allows intermediate pain measurements to vary as a function of prior perception and treatment.

3.6 Non-parametric Estimation

Suppose that the gabapentin trial investigator intended to estimate the perception effect, $\psi_p(\bar{P}_1 = \bar{1}, \bar{P}_2 = \bar{0}, A = 0) = E[Y_{0,\bar{1}}] - E[Y_{0,\bar{0}}]$, non-parametrically based on the causal graph depicted in figure 3.3. One approach would be to compare the average pain in the strata of patients who received a placebo and experienced side effects to the average pain for the subgroup that received a placebo and did not experience any side effects. This comparison would be equivalent to the conditional mean difference $\hat{E}[L|A = 0, \bar{P} = \bar{1}] - \hat{E}[L|A = 0, \bar{P} = \bar{0}]$. Under randomization and no unmeasured confounding assumptions, the conditional expectations of the form $E[L|A = a, \bar{P} = \bar{p}]$ are equal to the marginal expectation $E\left[L_{A=0,\bar{P}=\bar{p}}\right]$. In other words, if the group who received placebo $a = 0$ and had their perception fixed at $\bar{0}$ was

similar to the population in every aspect except for their treatment and perception levels, we would be able to claim that their average pain would represent the population's average pain under a hypothetical experiment where everyone received placebo $a = 1$, and their perception was fixed at $\bar{p} = \bar{0}$.

It is simple to see that randomization to both treatment and perception removes confounding. However, in practice, we do not intervene on patients' perception and the effect of perception on pain may be confounded by baseline variables. An investigator who intends to estimate a perception effects non-parametrically would not only be obligated to stratify on presence of side effects, but would also needs to stratify on baseline variables which may affect both perception and the outcome. Such stratification may produce many strata with sparse data depending on the size of the study, and the investigator would be faced with the curse of dimensionality. As the number of confounding factors increases, the investigator necessarily needs to rely on model assumptions.
On the other hand, the direct effect of the treatment on the outcome can still be estimated non-parametrically by excluding patients who experienced any side effects (stratifying on perception) and comparing the average pain score of those who received treatment to those who received placebo. Under this scenario, we would obtain an estimate for the conditional mean difference $E[L|A = 1, \bar{P} = \bar{0}] - E[L|A = 0, \bar{p} = \bar{0}]$ which under the randomization assumption for treatment is equal to the marginal mean difference $E\left[L_{1,\bar{0}}\right] - E\left[L_{0,\bar{0}}\right]$.

The investigator may be inclined to ignore side effects and proceed by an intent-to-treat analysis and interpret the results as the total effect of treatment. This might be justified by claiming that a portion of the ITT effect is due to the direct effect of the treatment, and the rest is due to an indirect effect of the treatment through perception. However the treatment effect obtained by ITT is not to be trusted since it captures a combination of the direct effect of the treatment and a confounded indirect effect through perception.

3.7 Semi-Parametric Estimation Methods

Based on the time ordering and causal graph assumptions of the previous section, the likelihood of the observed data may be written as $p(O) = \prod_j p(N(j)|Pa(N(j)))$, where $N(j)$ denote the nodes (i.e. observed variables), and $Pa(N(j)$ denote the parent nodes. (see Pearl[17], and van der Laan[38]) This probability distribution can further be represented as a product of two factors $p(O) = Qg$, where Q is the product of the conditional distribution of the non-intervention nodes and identifies the G-computation formula, and g represents the product of the conditional distribution of the intervention nodes and is often referred to as the treatment or censoring mechanism.[38]

Various semi-parametric estimators have been suggested for estimation of causal effects under time dependent interventions, including IPTW[20,21], Augmented IPTW[22,23,24], Maximum Likelihood Estimators[42], and Targeted Maximum Likelihood Estimators[38]. The IPTW and the augmented IPTW estimators belong to the category of estimating equation methodology[28]. The augmented IPTW estimator is defined as the solution of an estimating equation for the parameter of interest which is derived from the efficient influence function.[28] MLE is a plug-in estimator that estimates the distribution of the data and evaluates the parameter of interest using this distribution. MLE balances bias and variance with respect to the distribution of the data and not the parameter of interest. On the other hand, TMLE is a two stage estimators which improve on the MLE by minimizing the bias with respect to the parameter of interest by consistently estimating the treatment mechanism.[38] In this section we will reintroduce the MLE and TMLE estimators for the parameter of interest for the gabapentin trial in the longitudinal setting.

3.8 Maximum Likelihood Estimation (G-computation)

Consider the general observed data structure for the gabapentin trial:
$O = (L(0), A(0), P(1), L(1), P(2), L(2), \dots, P(K), L(K), P(K+1), Y = L(K+1))$
Assuming a causal graph similar to figure 3.3 generalized to multiple time points, the likelihood for the observed data can be factorized as:

$$p(O) = \prod_{j=0}^{K+1} p\big(L(j)\big|\bar{P}(j), \bar{L}(j-1), A(0)\big) \prod_{j=0}^{K+1} p\big(P(j)\big|\bar{L}(j-1), \bar{P}(j-1), A(0)\big) \times\ p\big(A(0)\big)$$

Where for $j = 0$, $\bar{L}(j-1)$ is empty, and $\bar{P}(j-1)$ is empty for $j = 0, 1$. The factors of the form $p(L(j)|\bar{P}(j), \bar{L}(j-1), A(0))$ will be denoted by $Q_{L(j)}$ and the other factors form the treatment mechanism. Note that the treatment mechanism consists of factors of the form $p(P(j)|\bar{L}(j-1), \bar{P}(j-1), A(0))$ which correspond to the time dependent perception variable and we denote by $g_{P(j)}$, and also $p(A(0))$ which corresponds to the static treatment assignment and we denote by $g_{A(0)}$.

Suppose we are interested in estimating the effect of the treatment had everyone's perception remained at no knowledge of the treatment, i.e. $\psi_a(\bar{P}(t) = \bar{p}) = \psi_{a=1}(\bar{P}(t) = \bar{0}) = E\big[L_{a=1,\bar{p}=\bar{0}}\big] - E[L_{a=0,\bar{p}=\bar{0}}]$, where $\bar{p} = \bar{0}$ means $P(t) = 0$ for all t. To estimate this effect we would be hypothetically intervening on $A(0)$, and $P(j)$ for all j. Let our treatment rule be denoted by $d(A = a, \bar{P} = \bar{p})$. The G-computation formula for the counterfactual distributions of the data under these hypothetical interventions, $d(A = a, \bar{P} = \bar{p})$, would be given by:

$$p_{d(a,\bar{p})}(O) = \prod_{j=0}^{K+1} p(L(j)|\bar{P}(j), \bar{L}(j-1), A(0))$$

Each of the counterfactual means can be estimated using a Monte Carlo simulation, by sequentially simulating from the conditional distributions $Q_{L(0)}, Q_{L(1)}, \dots Q_{L(K+1)}$ under the corresponding treatment rule.[38]

Given an estimator Q_n for Q, we obtain a substitution estimator for ψ:

$$\hat{\psi}_{a=1}(\bar{P}(t) = \bar{0}) = \hat{E}\left[L_{a=1,\bar{p}=\bar{0}}\right] - \hat{E}[L_{a=0,\bar{p}=\bar{0}}]$$

For the simple case of three time points, we have the following time ordering:

$$L(0), A(0), P(1), L(1), P(2), L(2),$$

and the likelihood can be factorized as:

$$p(O) = p\big(L(0)\big)p\big(L(1)\big|P(0), L(0), A(0)\big)p\big(L(2)\big|P(2), P(1), L(1), L(0)A(0)\big) \times$$
$$p\big(P(1)\big|L(0), A(0)\big)p\big(P(2)\big|L(1), L(0), P(1), A(0)\big) \times \ p\big(A(0)\big) = Q_0 Q_1 Q_2 g$$

As mentioned earlier, the G-computation formula involves modeling the Q part of the likelihood; possibly, using data adaptive loss based learning methods. To estimate a parameter such as $E\left[L_{a=1,\bar{p}=\bar{0}}\right] - E[L_{a=0,\bar{p}=\bar{0}}]$, each of the two means can be estimated separately. For instance, to estimate $\hat{E}\left[L_{a=1,\bar{p}=\bar{0}}\right]$, one would:

I) Obtain estimates $\hat{Q}\big(L(1)\big|P(0), L(0), A(0)\big)$, and $\hat{Q}\big(L(2)\big|P(2), P(1), L(1), L(0), A(0)\big)$

II) Generate a large dataset from the empirical distribution of $\hat{Q}\big(L(0)\big)$, plug in the data into the model for $L(1)$, fixing $A(0) = 1$, and $P(1) = 0$ for everyone to generate $L(1)$'s,

III) Plug in the generated $L(1)$'s and the other covariates into the model for $L(2)$, fixing $P(2) = 0$ for everyone

IV) Take the empirical mean $\frac{1}{n}\sum_{i=1}^{n} \hat{Q}\left(L(2)\middle|P(2) = 0, P(1) = 0, A(0) = 1, L_{A(0)=0,P(1)=0}(1), L(0)\right)$

3.9 Review of Semi-Parametric Efficient Estimation Theory

Here, we briefly deviate from our main discussion to introduce semi-parametric estimation theory without delving into much technical details. Our summary follows a much more detailed discussion of the concepts introduced by Bickel et al.[47] and Tsiatis[48]. In the next section, we will build upon the concepts introduced in this section to demonstrate the derivation of TMLE estimator (as in van der Laan[38]).

Even though the investigators collect data on many covariates in a study, the interest often lies in a low dimensional parameter,β, of the full data distribution. In such

setting, the use of semi-parametric models for the full data distribution which do not model redundant components are advantages since they lack biases due to misspecifications of the functional form of a parametric model. Suppose $Z_1, \dots, Z_n$ are identically an independently distributed (iid) with density belonging to a probability model (or class of densities) which might have generated the data. For instance, let $Z_i = (W_i, A_i(0), P_i(1), L_i(1), P_i(2), Y_i = L_i(2))$ denote the random vector for a single observation in the gabapentin trial. The densities in a model may be identified through a set of parameters θ. In cases where models may be described through a finite number of parameters, they are referred to as finite-dimensional parametric models. In other cases, the class of densities may be so large that the parameter θ is infinite dimensional. For such cases, we may be interested in β, a finite dimensional subset of θ. Hence, θ may be partitioned as (β, η), where β is the q-dimensional parameter of interest and η is the infinite dimensional nuisance parameter.[48]

Considering an underlying probability space for $Z_1, \dots, Z_n$, let $\mathcal{H}: \mathfrak{L} \to \mathbb{R}^q$ (where $\mathfrak{L}$ is the sample space) be the space consisting of q-dimensional random functions of Z, where $h(Z)$ has mean zero ($E[h(Z)] = 0$), and finite second moment ($E[h^T(Z)h(Z)] < \infty$). The space of all h that satisfy the above conditions is a linear space. In other words, if h_1, h_2 are elements of this space, for any real constants a and b, $ah_1 + bh_2$ also belongs to this space. The linear vector space of q-dimensional random functions with mean zero and finite second moment can be extended to a Hilbert space by defining an inner product $\langle h_1, h_2 \rangle$. For $h_1, h_2 \in \mathcal{H}$, let $\langle h_1, h_2 \rangle = E(h_1^T h_2)$, which is referred to as the "covariance inner product". Once an inner product has been defined, the norm of any element of $\mathcal{H}$ can be defined as $\|h\| = \langle h, h \rangle^{1/2}$. In addition, two vectors h_1, h_2 are called orthogonal if $\langle h_1, h_2 \rangle = 0$.[48]

A space $\mathcal{U} \subset \mathcal{H}$ is a linear subspace if $u_1, u_2 \in \mathcal{U}$ implies that $au_1 + bu_2 \in \mathcal{U}$ for scalars a, b. For Hilbert spaces, we have the following theorem for projection of an element of $\mathcal{H}$ onto a subspace $\mathcal{U}$:

Projection Theorem: Let $\mathcal{H}$ be a Hilbert space and $\mathcal{U}$ a linear subspace that is closed (i.e. contains all its limit points). For every $h \in \mathcal{H}$, there exists a unique $u_0 \in \mathcal{U}$ that is closest to h, or

$$\|h - u_0\| \leq \|h - u\| \text{ for all } u \in \mathcal{U}$$

Furthermore, $h - u_0$ is orthogonal to $\mathcal{U}$; that is $\langle h - u_0, u \rangle = 0$ for all $u \in \mathcal{U}$. u_0 is the projection of h onto the space $\mathcal{U}$, and is denoted by $\prod(h|\mathcal{U})$.[48] (See Tsiatis[48] for proof)

Most reasonable estimators for the parameter β are asymptotically linear and can be uniquely characterized by an influence function. An estimator $\hat{\beta}_n$ of β is said to be asymptotically linear if there exists a random vector (i.e. a q-dimensional random function) $\varphi^{q \times 1}(Z)$ such that $E[\varphi(Z)] = 0^{q \times 1}$, and $n^{1/2}(\hat{\beta}_n - \beta_0) = n^{-1/2} \sum_{i=1}^{n} \varphi(Z_i) + o_p(1)$.[48] Where $o_p(1)$ is a term that converges in probability to 0 as n goes to infinity, β_0 is the truth, and $E(\varphi\varphi^T)$ is finite and

nonsingular. The random vector $\varphi(Z_i)$ is referred to as the influence function of the estimator $\hat{\beta}_n$(Tsiatis).[48]

By the Central Limit Theorem, regular asymptotically linear estimators are asymptotically normally distributed; i.e.

$$n^{\frac{1}{2}}(\hat{\beta}_n - \beta_0) \xrightarrow{\mathcal{D}} N(0, E(\varphi\varphi^T))$$

For a single observation Z in a parametric model, where $Z \sim p_Z(z,\theta), \theta = (\beta^T, \eta^T)$, the score vector $S_\theta(z,\theta_0)$ is defined as the p-dimensional vector of derivatives of the log likelihood with respect to the element of the parameter θ, where $\theta_0$denotes the value of θ that generates the data.[48]

$$S_\theta(z,\theta_0) = \frac{\partial log p_Z(z,\theta)}{\partial\theta}\Big|_{\theta=\theta_0}$$

This vector can be partitioned according to the parameters of interest β and the nuisance parameters η as $S_\theta(Z,\theta_0) = \{S_\beta^T(Z,\theta_0), S_\eta^T(Z,\theta_0)\}^T$, where $S_\beta(Z,\theta_0) = \frac{\partial log p_Z(z,\theta)}{\partial\beta}\Big|_{\theta=\theta_0}^{q\times 1}$ and $S_\eta(Z,\theta_0) = \frac{\partial log p_Z(z,\theta)}{\partial\eta}\Big|_{\theta=\theta_0}^{r\times 1}$. [48] Under suitable regularity conditions the score vector $S_\theta(Z,\theta_0)$ has mean zero (i.e. $E[S_\theta(Z,\theta_0)] = 0^{p\times}1$). [48]

A tangent space can be defined as the finite dimensional linear subspace $\mathcal{T} \subset \mathcal{H}$ spanned by the p-dimensional score vector $S_\theta(Z,\theta_0)$ as the set of all q-dimensional mean-zero random vectors consisting of $B^{q\times p}S_\theta(Z,\theta_0)$ for all $q \times p$ matrices B. [48] For a parametric model, the subspace spanned by the nuisance score vector $S_\eta(Z,\theta_0)$ is referred to as the nuisance tangent space, Λ.[48]

To work with the nuisance tangent space for semi-parametric models, we simplify things by considering a finite-dimensional parametric sub-model contained within the semi-parametric model. For a semi-parametric model, the nuisance tangent space is defined as the mean-square closure (see Tsiatis[48] for the technical definition) of parametric submodel nuisance tangent spaces. Furthermore, the tangent space $\mathcal{T}$ can be decomposed as the direct sum of $\mathcal{T}_\beta$ and Λ, i.e. $\mathcal{T} = \mathcal{T}_\beta \oplus \Lambda$, where $\mathcal{T}_\beta$ is the subspace spanned by the score vector $S_\beta(Z,\theta_0)$. [38,48] In the section we will use the above setting and the projection theorem to derive the TMLE (following van der Laan[38]) by using the efficient influence function for the IPTW estimator.

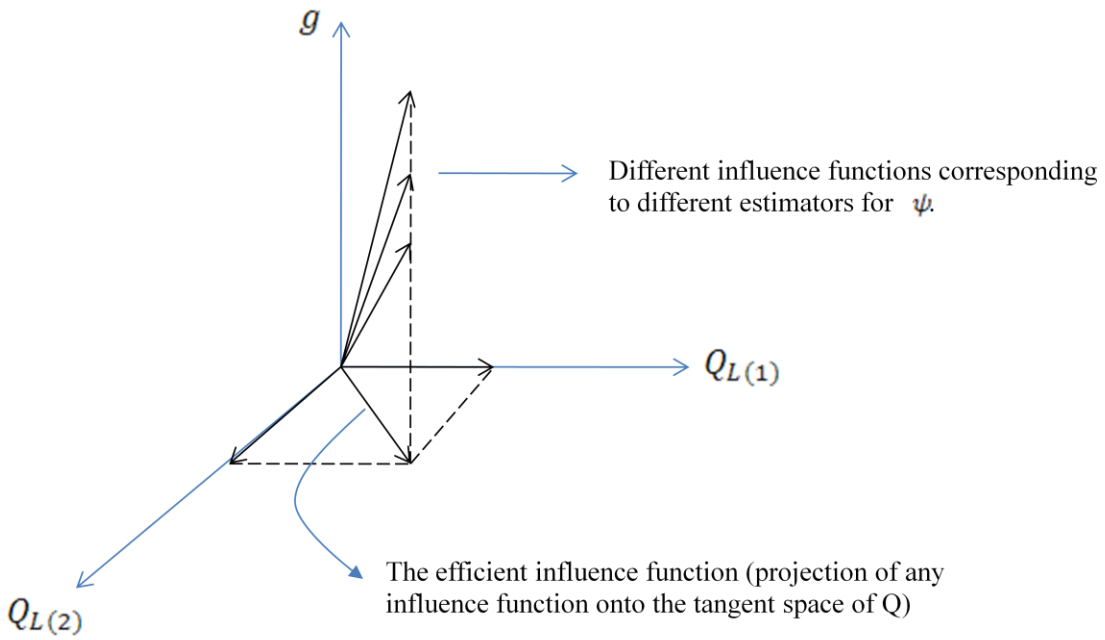
3.10 Targeted Maximum Likelihood Estimation (TMLE)

As discussed in chapter 2, TMLE is a two-stage estimator that modifies the Maximum Likelihood Estimator in a manner which reduces bias for the target parameter of interest if the treatment/censoring mechanism can be estimated consistently.[25,38] The first stage of TMLE involves defining an initial estimate for Q as in Maximum Likelihood estimation. This initial estimate can be obtained using data adaptive loss

based learning methods which will be discussed shortly. The second stage involves fluctuating the initial estimate for Q in a way which reduces bias for the parameter of interest.[38] The source of bias in the MLE of the parameter of interest is a result of the MLE being a plug-in estimator. For Maximum Likelihood estimation, the goal is optimal estimation of Q rather than the parameter of interest and thus bias and variance is balanced with respect to the density. However, when using TMLE one is interested in minimizing the bias for the parameter of interest. The fluctuated function for targeted estimation of the parameter of interest is the least favorable parametric submodel through Q.[38] This least favorable model is a model that has a score at zero fluctuation equal to the efficient influence function (canonical gradient of the pathwise derivative) of the target parameter ψ. [38]

In order to present the TMLE, we need to define an efficient influence function. Consider various asymptotically linear estimators β_n for a parameter β. Each of these estimators can be uniquely identified by an influence function. However, there is a unique influence function with the smallest variance (which attains the semi-parametric efficiency bound) and is referred to as the efficient influence function, $\varphi^*(Z_i)$.[48] We noted that the tangent space can be decomposed as $\mathcal{T} = \mathcal{T}_\beta \oplus \Lambda$. The likelihood of the observed data was factorized as $l(O) = \prod_j Q_{L(j)} \prod_j g_{P(j)} g_{A(0)} = Q.g$, and we estimated the parameters of Q. These parameters of interest for a parametric submodel of Q correspond to β, and the scores of Q span $\mathcal{T}_\beta$. In addition, g can be viewed as the nuisance tangent space, and the scores of g span Λ. From this point on we may use $\mathcal{T}_Q$ and $\mathcal{T}_\beta$ interchangeably.
The efficient influence function for a given parameter of interest may be constructed by projecting the influence function of any consistent asymptotically linear estimator for that parameter onto the tangent space $\mathcal{T}_Q$, where $\mathcal{T}_Q$ is the space spanned by the scores of a parametric submodel for p_0. For instance, to obtain the efficient influence function for $\psi = EY_d$ we may project the influence function for the IPTW estimator, $\varphi_{IPTW}(O) = \frac{I(A=a,\bar{P}=\bar{p})}{g(A,\bar{P}|W)} Y - \psi$, onto the tangent space of $Q(\mathcal{T}_Q)$. $\mathcal{T}_Q$, can be further decomposed as $\mathcal{T}_Q = \sum_j \mathcal{T}_{Q_{L(j)}}$, where the summation represents direct sum of the orthogonal subspaces. Here, the $Q_{L(j)}$ factors consist of functions of $L(j), Pa(L(j))$, with conditional mean zero, given the parents $Pa(L(j))$ of $L(j)$, for all j. Hence, to project an influence function onto the tangent space of Q, it suffices to project it onto each of the subspaces $Q_{L(j)}$.[38]

Figure 3.4 Efficient influence function: projection of an influence function onto the tangent space



Let the projection of the influence function onto $Q_{L(j)}$ be denoted by $\varphi_j^* = \prod(\varphi^*|\mathcal{T}_{Q_{L(j)}})$, $j = 0,1,2$.

Then

$$\varphi_0^*(O) = E[Y_d|L(0)] - \psi$$

$$\varphi_1^*(O) = \frac{I(A(0) = a, P(1) = p_1)}{g(A, P(1)|W)}\{E(Y_d|L(0), A(0), P(1), L(1) = 1) - E(Y_d|L(0), A(0), P(1), L(1) = 0)\} \times \{L(1) - E(L(1)|L(0), A(0), P(1))\}$$

$$\varphi_2^*(O) = \frac{I(A(0) = a, \bar{P} = \bar{p})}{g(A, \bar{P}|W)}\{L(2) - E\big(L(2)\big|L(0), A(0), P(1), L(1), P(2)\big)\}$$

The Targeted Maximum Likelihood Estimator can be defined as follows. Suppose we have an initial estimator $Q_{L(j)_n}$ for each $Q_{L(j)}$, $j = 0,1,2$. The marginal probability $Q_{L(0)}$ can be estimated using the empirical distribution of $L_i(0)$, $i = 1, \dots, n$. Conditional distributions for pain at the end of week 4, $L(1)$, and the final pain score $L(2)$ can be estimated using machine learning algorithms. After obtaining initial estimates for $Q_{L(j)}$, we fluctuate the initial estimates in the following manner:[38]

$$Q_n = (Q_{L(1)_n}, Q_{L(2)_n})$$

$$Q_{L(1)_n}(\varepsilon) = Q_{L(1)_n} + \varepsilon_1 C_{L(1)}(Q_n, g_n)$$

$$Q_{L(2)_n}(\varepsilon) = Q_{L(2)_n} + \varepsilon_2 C_{L(2)}(Q_n, g_n)$$

Where,

$$C_{L(1)}(Q_n, g_n) \equiv \frac{I(A(0)=a, P(1)=p_1)}{g(A, P(1)|W)}\{E(Y_d|L(0), A(0), P(1), L(1)=1) - E(Y_d|L(0), A(0), P(1), L(1)=0)\}$$

$$C_{L(2)}(Q_n, g_n) \equiv \frac{I(A(0)=a, \bar{P}=\bar{p})}{g(A, \bar{P}|W)}$$

The variables $C_{L(1)}(Q_n, g_n)$ and $C_{L(2)}(Q_n, g_n)$, are the coefficients of $L(j) - Q_{L(j)}$ from the projection of the IPTW influence function onto the tangent space, and are referred to as *clever covariates*.[38] Assuming that the Ys (or the errors) are normally distributed, we may use the normal densities as fluctuation models with mean $E_{Q_{Y_n}}\left(Y\middle|Pa(Y)\right) + \varepsilon_j C_{L(j)}(Q_n, g_n)$, $j = 1,2$, and constant variance σ^2.[38] In this case, the maximum likelihood estimator of ε_j is the least square estimator, and the score of ε_j at $\varepsilon_j = 0$ is equal to the $Q_{L(j)}$ component of the efficient influence function $C_{L(j)}(L(j) - E_Q(L(j)|Pa(L(j))))$. We can either estimate a common ε using the MLE, $\varepsilon_n = argmax_\varepsilon \prod_{j=1}^{2}\prod_{i=1}^{n} Q_{L(j)_n}(\varepsilon)(O_i)$, or obtain separate estimates of ε_j for each factor j, and iterate until convergence is achieved.[38] If we use a separate $\varepsilon_{L(j)}$, and first carry out the TMLE update for $Q_{L(2)_n}$, and use this updated $Q^*_{L(2)_n}$ in the TMLE update for $Q^*_{L(1)_n}$, then the targeted MLE algorithm converges in two simple steps.[38]

A parameter of interest can be estimated in the exact fashion as the previous section, with the difference that we use the fluctuated densities Q^*s rather than the Qs. To estimate $\hat{E}\left[L_{a=1,\bar{p}=\bar{0}}\right]$, one would

I) Obtain estimates $\hat{Q}^*\left(L(1)\middle|P(0), L(0), A(0)\right)$, and $\hat{Q}^*\left(L(2)\middle|P(2), P(1), L(1), L(0), A(0)\right)$

II) Generate a large dataset from the empirical distribution of $\hat{Q}\left(L(0)\right)$, plug in the data into the model for $L(1)$, fixing $A(0) = 1$, and $P(1) = 0$ for everyone to generate $L(1)$'s

III) Plug in the generated $L(1)$'s and the other covariates into the model for $L(2)$, fixing $P(2) = 0$ for everyone

IV) Take the empirical mean $\frac{1}{n}\sum_{i=1}^{n} \hat{Q}^*\left(L(2)\middle|P(2)=0, P(1)=0, A(0)=1, L_{A(0)=0,P(1)=0}(1), L(0)\right)$

Consider the likelihood function for the gabapentin data.

$p(O) = \prod_{j=0}^{2} p(L(j)|\bar{P}(j), \bar{L}(j-1), A(0)) \prod_{j=1}^{2} p(P(j)|\bar{L}(j-1), \bar{P}(j-1), A(0)) \times$ *p(A0)=Q0Q1Q2g*

We need to model pain scores at two separate stages. In the full targeted maximum likelihood approach, we fluctuate the density for both Q_1 and Q_2 spaces in the direction that minimizes bias for the parameter of interest. An alternative approach would be to only fluctuate Q_2. This approach restricts our fluctuation moves to the space of Q_2 and reduces the bias in the parameter of interest incurred by optimizing an estimate of Q_2 rather than the parameter of interest. Note that some of the patients are lost to follow up and we may have less data to model Q_2, and thus balancing bias and variance with respect to the density may come at a higher price in terms of bias for the parameter of interest. This last-step TMLE will be consistent if either Qor g is consistently estimated. However this estimator is not efficient, similar to the full TMLE.[25]

To use this last-step targeted maximum likelihood estimator we add the following clever covariate to the model for Q_2

$$\frac{I(A = a,\ \bar{P} = \bar{p})}{g(A, \bar{P}|W)} = \frac{I(A = a, P_1 = p_1, P_2 = p_2)}{P(A = a|W)P(P_1 = p_1|A = a, W)P(P_2 = p_2|A = a, P_1 = p_1, W)}$$

If we allowed the patients' perception to change without any restriction over time, considering that treatment and the perception variables are all binary, there would be eight possible combinations of treatment and perception variables. However, we are assuming that once a patient observes a side effect, his perception switches on and does not change for the rest of the trial. This assumption, limits the number of possible treatment and perception combinations to six, since we never observe the sequences: $A(0) = 0, P(1) = 1, P(2) = 0$ or $A(0) = 1, P(1) = 1, P(2) = 0$. Thus, any probability distribution for the treatment mechanism would assign 0 probabilities to those sequences.

3.11 Estimation of the Gabapentin Trial Parameters

We started the analysis of the gabapentin trial by creating average pain scores for the screening week, week 4 and week 8 of the trial. If patients were missing pain scores for any days during a specific week, the rest of the pain scores during that week were used to calculate the average, and only if a patient was missing all the seven pain scores we considered his pain score for that particular week as missing. In addition to the pain variables, we created two variables for perception. The first variable was an indicator for presence of any treatment related side effects between randomization and the end of week 3, and the second variable was and indicator for presence of any treatment related side effects between randomization and the end of week 7. The set of baseline covariates used in our analysis included sex, race, height, weight, age, mean baseline pain score, and mean baseline sleep score. We demonstrate how the parameters of section 4 were estimated by focusing on only one of the parameters: the treatment effect had everyone's perception remained at no knowledge of the treatment, i.e. had no one experienced treatment related side effects: $\psi_a(\bar{P}(t) = \bar{0}) =$

$E\left[L_{1,\bar{0}}\right] - E[L_{0,\bar{0}}]$. Estimation of the other parameters may be carried out in a similar fashion. To estimate $\psi_a(\bar{P}(t) = \bar{0})$, one can estimate each of the two marginal means, $E\left[L_{1,\bar{0}}\right]$ and $E[L_{0,\bar{0}}]$, and take their difference.

3.11.1 Maximum Likelihood Estimation (MLE)

Simulation based estimation for our parameters of interest involve modeling intermediate and final pain scores based on the preceding parent nodes and sequentially generating from those models. For Maximum Likelihood Estimation, since we intervene on treatment and perception, we do not model the treatment mechanism and solely rely on the models for pain scores. To model pain at the end of week 4 and the final pain at the end of week 8, we employed the DSA[32] algorithm. Using the DSA, we fit a logistic regression model for the intermediate pain $L(1)$, with treatment $A(0)$, perception $P(1)$, and the interaction between them forced into the model. To model pain at the end of the trial $Y = L(2)$, we used the DSA to fit a linear regression model with treatment $A(0)$, perception $P(2)$, and the interaction between them forced into the model. Perception $P(1)$ was not forced into the model for $L(2)$, however it was contained in the covariate space that the DSA algorithm searched for the best model. Both models for pain scores were restricted to a maximum size of 10 covariates, maximum order of 2, and possibility of two way interactions.

After modeling the pain scores, we created a new dataset of the baseline covariates by sampling the patients 100,000 times with replacement. Depending on which marginal mean, $E\left[L_{a,\bar{p}}\right]$, was being estimated, a column of 0's or 1's were added for treatment ($A = 0$ or $A = 1$), and another column of 0s or 1s were added for the patients' perception up to the end of week 3 ($P(1) = 0$ or $P(1) = 1$). The baseline covariates,W, treatment $A(0)$, and perception $P(1)$, were plugged into the logistic regression model for $L(1)$ to generate predicted probabilities for having a high pain score at the end of week 4. Binary $L(1)$ values were obtained by generating from a Bernoulli distribution with the predicted probability for each observation. The column of $L(1)$ pain scores and an additional column of 0's or 1's for perception at the end of week 7 ($P(2) = 0$ or $P(2) = 1$) were added to the generated dataset. Finally, the variables in the generated dataset were plugged into the model for $L(2)$ to generate final pain scores. The mean of the predicted $L(2)$ pain scores was the final estimate for the marginal mean $\hat{E}\left[L_{a,\bar{p}}\right]$. Each of the parameters of interest was estimated by estimating the two corresponding marginal means and taking the difference.

Bootstrap standard errors were estimated by re-sampling the original data 2000 times with replacement. For each bootstrap sample, we performed model selection using DSA, sequentially generated from the models to construct a simulated dataset, and finally, estimated the desired parameters as above. Once again, visual checks on the bootstrap distributions showed symmetric distributions of the bootstrap estimates

around the full-data estimate. For each parameter, the estimated standard error was used to calculate a two-sided Wald test statistic and a subsequent p-value.

3.11.2 Last-Step Targeted Maximum Likelihood Estimation

Targeted Maximum Likelihood Estimation requires updating the pain models by a clever covariate which depends on the probability of receiving a certain treatment and perception combination. The last-step TMLE involves only updating the pain model for week 8 by a given clever covariate while leaving the model for pain at the end week 4 unchanged. The clever covariate is a function of the treatment mechanism and thus requires modeling perception nodes $P(1)$ and $P(2)$. For the last-step TMLE the clever covariate is given by:

$$C_{L(2)}(Q_n, g_n) = \frac{I(A(0) = a, \bar{P} = \bar{p})}{g(A, \bar{P}|W)} = \frac{I(A(0) = a, P(1) = p_1, P(2) = p_2)}{P(A(0) = a, P(1) = p_1, P(2) = p_2|W)}$$
$$= \frac{I(A(0) = a, P(1) = p_1, P(2) = p_2)}{P(A(0) = a)P(P(1) = p_1|A(0), W)P(P(2) = p_2|A(0), P(1), W)}$$

Since treatment is assigned at random, the randomization probability (0.5) can be used for $P(A(0) = a)$. However, van der Laan[25] shows that using the empirical proportions of treated patients results in a gain in efficiency. Thus, we estimated $P(A(0) = a)$ by the proportion of the patients who received treatment.

Two separate logistic regression models were fit using the DSA algorithm for perception$P(1)$, and $P(2)$. The models for perception were restricted to a maximum size of 10 covariates, maximum order of 2, and possibility of two way interactions. Probability of receiving a specific combination of treatment and perception was calculated based on the empirical probability of receiving treatment and the two logistic regression models for perception. Since perception works as a switch, and a perception $P(1) = 1$ forces perception $P(2)$ to be equal to 1, we manually set $P(A(0) = a)P(P(1) = 1|A(0), W)P(P(2) = 1|A(0), P(1), W)$ to $P(A(0) = aPP1=1A0,W\times1$.

Once the $C_{L(2)}$ term is approximated, the coefficient ε_2 can be estimated using maximum likelihood for the regression model $L(2)$ assuming Gaussian error. However, if the probabilities of receiving some combinations of treatment and perception are small for some strata of the baseline variables, the Experimental Treatment Assumption will be violated. In this scenario, the values of the clever covariate can be large which may result in out of range predicted pain scores and thus inflation of the variance. An additional efficiency enhancement can be gained from constraining the model to properly predict the outcomes within the known limited range for the pain scores (0-10). Specifically, one can transform the dependent variable to lie between 0 and 1, and use a TMLE logistic regression approach that guarantees that all predicted scores fall within the known range.[34]

For this purpose, we transformed the predicted pain scores from the $L(2)$ model to values between 0 and 1 by subtracting the minimum predicted pain (which we call a) from each predicted value and dividing the values by the range of the predicted pain scores,$a - b$, where b is the maximum predicted pain score. The restricted pain scores were logit transformed and bounded away from 0 and 1 by setting logit values less than 0.01 equal to 0.01 and logit values greater than 0.99 equal to 0.99.[34] Subsequently, we fit a logistic regression model using the logit transformed values as the outcome, and $Q_{L(2)_n}$as a fixed offset to estimate the coefficient ε. The magnitude of the estimated ε depends on the amount of residual confounding (for estimation of the targeted mean parameter) along the direction of $C_{L(2)}(A, P, W)$. In our case, convergence is achieved in one step and there is no need for iteration: see Van der Laan[38]. The predicted probabilities from this model can be transformed back to the correct scale by multiplying each value by $a - b$ and adding a.

To estimate marginal means of interest, $E\left[L_{a,\bar{p}}\right]$, we proceeded as in MLE by sequentially generating data from $W, L(1)$, and the updated $L(2)$. Once again, bootstrap standard errors were estimated by re-sampling the observations with replacement 2000 times and for each bootstrap sample performing model selection, forming the clever covariate and updating the $L(2)$ model, generating a simulated dataset from the models, and finally, estimating the desired parameters by taking the mean of the $L(2)$ generated pain scores. For each parameter of interest, the estimated standard error was used to calculate a two-sided Wald test statistic and a subsequent p-value.

3.11.3 One-step Targeted Maximum Likelihood Estimation

The one-step TMLE involves updating both pain models for weeks 4 and 8. The clever covariate for the $L(2)$ pain model is the same as in one-step TMLE, and the clever covariate for $L(1)$ pain model is given by:

$$C_{L(1)}(Q_n, g_n) \equiv \frac{I(A(0) = a, P(1) = p_1)}{P(A(0) = a)P(P(1) = p_1|A(0), W)}\{E(Y_d|L(0), A(0), P(1), L(1) = 1) - E(Y_d|L(0), A(0), P(1), L(1) = 0)\}$$

Note that this clever covariate is dependent on the final pain score, Y, under a specific intervention d, and thus the parameter would be estimated by iteration. However, van der Laan[38] shows that by updating the final model ($L(2)$) first, and reversely using the predicted values from the final model in the previous models ($L(1)$) , TMLE converges in one step. To update the first pain model ($L(1)$) we estimated $\frac{I(A(0)=a,P(1)=p_1)}{P(A(0)=a)P(P(1)=p_1|A(0),W)}$ by using the empirical probability of receiving treatment for $P(A(0) = a)$, and the predicted probabilities from the perception model for $P(1)$. For each parameter of interest we estimated $E(Y_d|L(0), A(0), P(1), L(1) = 1)$, and $E(Y_d|L(0), A(0), P(1), L(1) = 0)$ by taking the mean of the predicted $L(2)$ pain scores from the last-step TMLE, once among the patients who had their intermediate

pain $L(1) = 1$ and another time among those who had $L(1) = 0$. After obtaining an estimate for the clever covariate, we fit a logistic regression model using the intermediate pain scores as the outcome, and $\hat{Q}_{L(1)_n}$as a fixed offset to estimate the coefficient for the clever covariate ε_1.

As for MLE, patients' baseline covariates were sampled 100000 times. The baseline covariates and the appropriate values of treatment $A(0)$ and perception $P(1)$ (depending on the marginal mean being estimated) were plugged into the updated model for $L(1)$ to obtain probabilities for having a high pain at the end of the 4th week. Binary pain values were generated based on these probabilities by sampling from a Bernoulli distribution for each replication. Baseline covariates, generated intermediate pain scores, and appropriate values of $A(0)$, $P(1)$, and $P(2)$ were plugged into the updated model for $L(2)$ to generate the final pain scores. The marginal mean of interest was obtained by taking the mean of final scores, and parameters of interest were estimated by taking the difference of the appropriate marginal means. The above process was repeated for 2000 different bootstrap samples to obtain standard errors for the parameters of interest.

3.12 Results for the Gabapentin Trial

Using an ITT analysis, Backonja et al.[11] had reported a significantly lower (1.2 points, p-value < 0.001) average pain score for the treatment group compared to the placebo group. Additionally, once patients who reported dizziness were excluded, the mean difference in pain scores between the treatment and the placebo groups remained almost the same (1.19 points, p-value=0.002). After excluding patients who experienced somnolence, even though the magnitude of the treatment effect decreased to 0.81, the difference remained significant (p-value= 0.03). In our cross-sectional analysis we saw that the treatment did not have any effect on pain, had the patients perception regarding the treatment remained at no knowledge of the treatment (i.e. no side effects) (0.78 points, p-value=0.18). However, the treatment would have had a significant effect on the outcome had the patients' perception been fixed at believing they were on treatment (i.e. side effects present)(1.98 points, p-value=0.04). In addition, we observed that the perception effect, had everyone received placebo, would not have been significant (-0.07 points, p-value=0.93), but had everyone received the active treatment, the perception effect would have been borderline significant (1.12 points, p-value=0.07). Overall, the results portrayed an interaction between perception and treatment.

Our time dependent analysis yields similar results to the cross-sectional setting with the difference that standard errors are smaller and there seems to be a gain in efficiency. Table 3.1 shows the results of MLE estimates for four different comparisons: 1) the treatment effect had everyone remained at no knowledge of the treatment or believed to be on placebo throughout the trial (i.e. no side effects); 2) the treatment effect had everyone believed to be on treatment throughout the trial (i.e.

side effects during the titration period); 3) the perception effect had everyone received placebo; 4) finally, the perception effect had everyone received treatment.

Table 3.1 Estimated longitudinal parameters using MLE and the corresponding 95% confidence intervals

Parameter (MLE 2000 B-samples)	G-comp	P-value	95 % CI
$\psi_a(\bar{P}(t)=\bar{0}) = E[Y_{0,\bar{0}}] - E[Y_{1,\bar{0}}]$	0.22 (0.41)	0.59	(-0.58, 1.02)
$\psi_a(\bar{P}(t)=\bar{1}) = E[Y_{0,\bar{1}}] - E[Y_{1,\bar{1}}]$	2.18(0.76)	0.004	(0.69, 3.67)
$\psi_p(\bar{P}_1=\bar{0},\bar{P}_2=\bar{1},A=0) = E[Y_{0,\bar{0}}] - E[Y_{0,\bar{1}}]$	-0.58 (0.75)	0.44	(-2.05, 0.89)
$\psi_p(\bar{P}_1=\bar{0},\bar{P}_2=\bar{1},A=1) = E[Y_{1,\bar{0}}] - E[Y_{1,\bar{1}}]$	1.38 (0.53)	0.009	(0.34, 2.41)

According to this table, had everyone remained at no knowledge of the treatment or believed to be on placebo, the treatment would have reduced the average pain by a magnitude of 0.22 points (on a 0-10 scale) with a 95% confidence interval of (-0.58,1.02). In other words, had no one experienced any side effects, the treatment would not have been statistically significantly different from 0. A similar conclusion can be drawn based on the TMLE estimates for $\psi_a(\bar{P}(t)=\bar{0})$. Tables 3.2 and 3.3 show the last-step and the one-step TMLE estimates for the treatment and perception effects of interest. Since the results from the last-step TMLE and the one-step TMLE are similar, we focus on comparing the MLE estimates with the one-step TMLE here and come back to the comparison of the two different types at the end of the section. Based on table 3.3, had no one experienced any treatment related side effects, the treatment effect would have been a reduction of 0.55 points with a 95% confidence interval of (-0.90, 2.00). Even though the magnitude of the estimate is more than twice the MLE estimate, the treatment effect is still statistically insignificant.

Table 3.2 Estimated longitudinal parameters using last-step TMLE and the corresponding 95% confidence intervals

Parameter (last-step TMLE 2000 B-	TMLE	P-value	95 % CI
$\psi_a(\bar{P}(t)=\bar{0}) = E[Y_{0,\bar{0}}] - E[Y_{1,\bar{0}}]$	0.61 (0.65)	0.35	(-0.66, 1.88)
$\psi_a(\bar{P}(t)=\bar{1}) = E[Y_{0,\bar{1}}] - E[Y_{1,\bar{1}}]$	2.77 (1.23)	0.02	(0.36, 5.18)
$\psi_p(\bar{P}_1=\bar{0},\bar{P}_2=\bar{1},A=0) = E[Y_{0,\bar{0}}] - E[Y_{0,\bar{1}}]$	-0.19 (1.22)	0.87	(-2.58, 2.20)
$\psi_p(\bar{P}_1=\bar{0},\bar{P}_2=\bar{1},A=1) = E[Y_{1,\bar{0}}] - E[Y_{1,\bar{1}}]$	1.96 (0.78)	0.01	(0.43, 3.48)

On the other hand, had everyone experienced a side effect during the titration period (had been unmasked for the duration of the trial), the treatment effect would have been statistically significant. The MLE estimate for this treatment effect is a reduction of 1.38 points with a 95% confidence interval of (0.34, 2.41). The less biased one-step TMLE estimate for this treatment effect is a drop of 2.72 points with a 95% confidence interval of (0.38, 5.16). The difference between the treatment effect had everyone seen a side effect versus had no one seen a side effect, suggest an interaction between the treatment and presence of side effects (or an interaction between the treatment and patients' perception).

Table 3.3 Estimated longitudinal parameters using full TMLE and the corresponding 95% confidence intervals

Parameter (Full TMLE 2000 B-	TMLE	P-value	95 % CI
$\psi_a(\bar{P}(t)=\bar{0}) = E[Y_{0,\bar{0}}] - E[Y_{1,\bar{0}}]$	0.55 (0.74)	0.46	(-0.90, 2.00)
$\psi_a(\bar{P}(t)=\bar{1}) = E[Y_{0,\bar{1}}] - E[Y_{1,\bar{1}}]$	2.72(1.22)	0.02	(0.38, 5.16)
$\psi_p(\bar{P}_1=\bar{0}, \bar{P}_2=\bar{1}, A=0) = E[Y_{0,\bar{0}}] - E[Y_{0,\bar{1}}]$	-0.19 (1.21)	0.86	(-2.56, 2.18)
$\psi_p(\bar{P}_1=\bar{0}, \bar{P}_2=\bar{1}, A=1) = E[Y_{1,\bar{0}}] - E[Y_{1,\bar{1}}]$	1.97(0.83)	0.02	(0.34,3.60)

In addition to treatment effects of interest, we estimated two different perception effects $\psi_p(\bar{P}_1=\bar{0}, \bar{P}_2=\bar{1}, A=0)$ and $\psi_p(\bar{P}_1=\bar{0}, \bar{P}_2=\bar{1}, A=1)$. Based on the MLE results, had everyone received no treatment and their perception fixed at no knowledge of the treatment or placebo throughout the trial, versus their perception fixed at treatment throughout the trial, their average pain score would have been 0.58 points lower with a 95% confidence interval of (-2.05, 0.89). Once again, the magnitude of this effect is not statistically different from 0. TMLE estimates the magnitude of this effect to be closer to 0, at a 0.19 point reduction of pain.
On the other hand, had everyone received treatment and their perception fixed at treatment throughout the trial versus no knowledge of the treatment or placebo throughout the trial, their pain would have been significantly lower. MLE estimate this reduction to be 1.38 points with a 95% confidence interval of (0.34, 2.41), and TMLE estimates the reduction to be a bit higher at 1.97 points with a 95% confidence interval of (0.34,3.60).

3.13 Discussion

In this chapter, we examined perception parameters comparable to those obtained in the second chapter. However, it is possible, and it may be interesting, to estimate the

effect of other perception patterns on the pain outcome as well. For instance, for the gabapentin trial, it is possible to estimate $\psi_p(\bar{P}_1 = \bar{0}, \bar{P}_2 = (0,1), A = 0) = E[Y_{0,\bar{0}}] - E[Y_{0,(0,1)}]$, in which patients' perception remains at no knowledge of the treatment through the trial in one scenario, and in the second scenario their perception changes during the second half of the trial to believing that they are on treatment (i.e. observing a side effect in the second half of the trial).

In the cross-sectional analysis we noted that using treatment related side effects as a proxy for a patient's perception regarding his treatment results in an asymmetry in the analysis of the gabapentin trial. Our analysis does not account for the possibility of a patient's perception switching to placebo at a specific time point during the course of the trial. A patient may start believing that he is on a placebo due to lack of efficacy of the administered treatment, biasing the patient's subjective pain score, upward. Depending on the treatment arm which the patient belongs to, this bias can result in either an increase or decrease in the estimated treatment effect. Unfortunately, it is impossible to determine this bias in case of the gabapentin trial without additional information.

In addition to the asymmetry that arises as the result using side effects as a proxy for perception, it is also impossible to disentangle perception effect from a stronger treatment effect. It can be argued that the treatment is having a stronger effect on patients who experience treatment related side effects in the treatment arm. This phenomenon is referred to as Philip's paradox. [49] Due to this paradox, other authors have suggested that the treatment be titrated before the start of the actual trial. For the gabapentin trial, the argument for a stronger treatment effect may be more reasonable for patients who experience side effects earlier in the trial during the titration period as the investigators try to find the maximum tolerable dosage. Yet, any treatment related side effects in the placebo arm cannot be attributed to the active component and is likely due to the patient's perception regarding his treatment.

Issues with using side effects as a proxy for perception and Philip's paradox highlight the need for collecting data (possibly longitudinally) on patients' perception regarding their treatment in randomized clinical trials. Although, it can be argued that questioning patients on their perception regarding their treatment arm may affect their perception itself. For instance, patients who have no knowledge of their treatment assignment might reevaluate and change their perception if they are asked to identify their treatment group. In 2003, the Food and Drug Administration (FDA) noted that treatment related side effects have the potential to unmask subjects and investigators, and may bias subjective study end points. (Office of Therapeutics Research and Review, Center for Biologics Evaluation and Research, FDA 2003)[50] They recommended that a questionnaire be administered at the completion of the study to investigate the effectiveness of blinding of the subjects and the investigators.[50] In a recent home drinking water intervention trial for estimating rates of highly credible gastrointestinal illness, Colford et al.[51] assessed whether participants could be successfully blinded to a sham or active water treatment device installed underneath the kitchen sink. They administered a questionnaire every 2

weeks for a 4 month period, and the participants were asked to rate their degree of certainty regarding having the active device. Blinding of the participants was assessed using a blinding index, and the investigators concluded that the participants were successfully blinded to their treatment assignment.[51] James et al.[52], Howard et al.[53], and Bang et al.[54] have introduced different indices for the degree of blinding in clinical trials. Although these indices tell the investigators whether blinding has been effective or not, they do not directly explore the effect of unmasking on the outcome of interest.

In this chapter, we demonstrated the use of efficient semi-parametric estimation methods for estimating causal parameters in a longitudinal setting. We described how one can formulate a causal parameter of interest using the counterfactual framework and estimate the parameter using the MLE and TMLE estimators. In the next chapter, we use semi-parametric modeling in a much different context, where the primary goal of statistical modeling is prediction rather than explanation.

Chapter 4

Re-examining the Framingham Coronary Heart Disease Models

4.1 Background

Recent statistical history of observational epidemiology has been one dominated by the use of parametric statistical regression models. Statistical modeling is typically used for one of two purposes: explanation of a phenomenon or prediction. Although there is a longstanding tradition of parametric modeling for explanation and prediction in public health and medical research, the objectives of statistical modeling are often not a priori clarified by the investigators. Rarely have statistical models been applied to medical data uniquely for the purpose of prediction, and often the use of explanatory models has been extended to prediction. Ambiguities regarding the objectives of statistical modeling have led investigators to employ models for one purpose, despite the fact that such models are intended and developed for other purposes. For instance, to examine the form of models apparently designed for prediction to make statements regarding the relative influence of variables.

In addition, the class of models considered has been traditionally based more on convenience, rather than a scientifically rigorous decision (i.e. what is actually known) about the statistical model of the data-generating process.

In most health studies, there is little knowledge outside the data about the true functional form of the model. This lack of knowledge to constrain the model could have resulted in three different directions: 1) estimating aspects of the data-generating distribution such as associations by making as few modeling assumptions as possible, 2) imposing arbitrary models that are convenient for returning estimates of a parameter of interest such as adjusted associations, or 3) A combination of the last two that acknowledges the lack of knowledge about the statistical model but uses a low-dimensional approximation.[55]

The first requires either a (possibly coarse) discretization of the covariate space, or sophisticated methods of searching through a large model space. The second requires widely available regression packages made trivial to implement with modern computers. Evident from publications in medicine and public health, the latter approach has dominated the field.

A further objectionable aspect of common practice is that the algorithms used to choose such parametric models are often not pre-specified, and thus, the resulting estimates and inference have been the result of an ad hoc data adaptive procedure. Though this type of procedure can take the form of a reproducible algorithm (e.g., as

formal step wise regression with main terms regression), they are often a complicated process, which can involve the screening of covariates based on analyses such as contingency tables, and somewhat random processes such as dialogue among researches about the importance of a variable. Thus, there is a complicated feedback of model and data, a process many in good faith try to relate in publications, but is certainly almost always ignored in the reported inference. Inference based on estimates from such models can be wildly optimistic since it ignores the fact that the data has been used for both estimation and validation. Without a well-defined algorithm for deriving an estimate, one has no estimator, and thus, no way to define the statistical properties of the estimate. For instance, in many approaches, one cannot define what is being actually estimated, as that itself is random given the procedure.[56]

In this chapter we tackle the prediction and explanation objectives separately in re-analyzing the Framingham study. We present a roadmap (based on previously published work on estimation in semi-parametric models) that demonstrates how robust information can be derived by applying more rigorous methods, which allow more reliable statistical inference (e.g., p-values that are calculated based on experiments that are commiserate to what was actually done). In particular, we use semi-parametric loss based estimation methods for prediction, and estimate variable importance using TMLE for explanation. As described in chapter 3, semi-parametric methods lack biases due to misspecifications of the functional form of a parametric model. Furthermore, using machine learning methods reduces any biases inherent in ad hoc modeling approaches.

4.2 Framingham Coronary Heart Disease Risk Scores

Coronary heart disease has been the leading cause of death in the United States since 1921.[57] In 2006, cardiovascular disease was responsible for 31.7% of all deaths; 26.0% from heart disease and 5.7% from stroke.[57] Early prediction of CHD may allow changes to modifiable life-style factors such as diet, exercise, and smoking that may in turn lower the risk of CHD. Various risk prediction models have been developed for CHD including the Framingham models. In 1998, Wilson et al.[58] incorporated the Joint National Committee (JNC-V) blood pressure and National Cholesterol Education Program (NCEP) cholesterol categories into sex-specific coronary heart disease (CHD) risk prediction models using a sample from the Framingham Heart Study. The objectives of the study were defined as examining the association of JNC-V blood pressure and NCEP cholesterol categories with CHD, as well as developing a simple coronary disease prediction algorithm which allowed physicians to predict CHD risk in patients without overt CHD.[58]

The study was designed as a prospective, single-center study within the Framingham Heart Study. The sample included 2489 men and 2856 women 30 to 74 years old at the start of their follow-up in 1971 to 1974. Risk factors considered for the models included age, blood pressure, cigarette smoking, total cholesterol (TC), low-density

lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), and diabetes. Study subjects were followed up for 12 years, and the outcome was defined as occurrence of any "hard CHD" which included recognized and unrecognized myocardial infarction, coronary insufficiency, and death due to CHD. The relationship between various independent variables and CHD outcome were tested using age-adjusted Cox proportional hazard regression, and the discriminatory ability of the risk prediction models were evaluated by the accompanying c-statistics.

The models were fit once including TC categories and excluding LDL-C categories, and a second time including LDL-C categories and excluding TC categories.[58] Quadratic terms for age were considered in the risks models, and among women the term was found to be significant. Furthermore, Wilson et al. argued that the relative risk for TC and CHD declines with age, and thus interaction terms for TC and age were considered in the prediction models. However, neither the interaction between age and TC nor an interaction between age and LDL-C was found to be significant in either sex. The model building process used for development of the Framingham models involved a combination of investigators' knowledge regarding the association of the risk factors with the outcome and a set of assumptions regarding the type and form of the models. Parameters of interest in the Framingham study were the coefficients of the risk factors included in the Cox proportional hazard model.

Wilson et al. evaluated the prediction models in various ways. The efficacy of prediction with continuous variables was compared with that obtained with categorical variables using a c-statistic which equals to the area under receiver operating characteristic curve. Area under the curve (AUC) provides a measure of the discriminatory power of a prediction model. Using AUC as an evaluation criterion assumes that a higher AUC directly corresponds to a lower value of the risk. The investigators concluded that the curves were nearly identical for the continuous and categorical models, and that TC and LDL-C categories had similar effects.[58] The c-statistics associated with TC categories were 0.74 in men and 0.77 in women for continuous variables by proportional hazards or accelerated failure models, and 0.73 in men and 0.76 in women for categorical variables. Similarly, the corresponding c-statistics associated with LDL-C categories were 0.74 in men and 0.77 in women for continuous variables by proportional hazards or accelerated failure models, and 0.73 in men and 0.77 in women for categorical variables.[58]

To evaluate the relative importance of each variable in the Framingham models, the investigators explored the association of each risk factor with the outcome using the coefficients of the Cox proportional hazard models. Wilson et al. noted that the CHD rates were significantly associated with the specified categories of blood pressure, TC, HDL-C, and LDL-C in both sexes. The relatives risks obtained from the Framingham study are shown in table 4.1.

Table 4.1 Multivariable-Adjusted Relative Risks for CHD According to TC Categories

		Men		Women	
		Relative Risk	95% CI	Relative Risk	95% CI
Blood Pressure	Normal	1.00	Referent	1.00	Referent
	High Normal	1.31	0.98–1.76	1.30	0.86–1.98
	Hyper I	1.67	1.28–2.18	1.73	1.19–2.52
	Hyper II-IV	1.84	1.37–2.49	2.12	1.42–3.17
Smoke	Yes/No	1.68	1.37–2.06	1.47	1.12–1.94
Diabetes	Yes/No	1.50	1.06–2.13	1.77	1.16–2.69
TC (mg/dL)					
	<200	1.00	Referent	1.00	Referent
	200-239	1.31	1.01–1.68	1.51	1.01–2.24
	>=240	1.90	1.47–2.47	1.72	1.15–2.56
HDL-C (mg/dL)					
	<35	1.47	1.16–1.86	2.02	1.29–3.15
	35-59	1.00	Referent	1.00	Referent
	>=60	0.56	0.37–0.83	0.58	0.43–0.79

In addition to assumptions regarding the form of the model, Framingham models appear to have been estimated and validated on the same learning set. It is well-known that performance estimates based on the same data used for estimation can be optimistic. In this case, given that model selection appears ad hoc, and all of the data appears to have been used to select the model, we cannot derive an unbiased estimate of the risk for their approach. We can only obtain an unbiased estimate the risk for the final model, assuming (incorrectly) that it was chosen a priori.

4.3 The Estimation Roadmap

In this section, we present an alternative approach based on established theory of so-called loss-based estimation in a semi-parametric model. This approach provides a benchmark for estimator comparison (loss-based) and admits that we are typically ignorant of the underlying true model form (semi-parametric). To develop the loss-function necessary to compare competing models and estimate the relative optimal one, we begin by defining the parameter of interest explicitly as the minimizer of an expected loss.

In a series of papers[59,60] and summarized in a recent book[61], van der Laan et al. develop a unified loss-based cross-validation methodology for estimator construction, selection, and performance assessment in presence of censoring. Risk prediction modeling and performance assessment can be approached under the same framework. Some aspects of loss-based performance evaluation for risk prediction have received

attention in recent years. For instance, Gail and Pfifer[62] review various risk prediction performance criteria under a decision theoretic framework and develop specific loss-function based criteria for different clinical applications. Yet, it is rare to find studies which incorporate these loss functions into estimation of a risk prediction model and more so to use such criteria to choose among a large class of competitors.

Van der Laan and Rose[61] proposed a general roadmap for estimation which can be partially summarized in the following steps: 1) define the research question of interest, 2) define the parameter as the minimizer of an expected loss, or risk, apart from any specific statistical model of the data, 3) estimate the relevant components of the data-generating distributions using an ensemble learning (Super learner) approach, using pre-specified algorithms only constrained by the assumptions about the model that are known to be true, and finally, 4) draw conclusions based on robust sampling-based inference. This above roadmap will guide us in creating a risk prediction model for CHD.

For the prediction goal, we are specifically interested in an individual's risk of CHD given his characteristics. Suppose we observe $O_i = (Y_i, W_i) \sim P_0, i = 1, \dots, n$ where Y_i is the outcome, W_i are the predictors, and P_0 is the unknown underlying data – generating distribution. Specifically, let the true parameter of interest, $\psi = \psi(P_0)$, be defined as the minimizer of the mean loss function, $E[L(O, \psi)]$ over the entire set of possible models for ψ, from a class of $\mathcal{M}$ (in our case this will be semi-parametric meaning for now, nearly all models) for the data-generating distribution, P_0. It should be noted that the syntax $\psi(P_0)$ is used to recognize that the parameter of interest can be understood as a mapping applied to the data-generating distribution. Our goal is to estimate a model for P_0, but a model that does the "best" job of estimating, not the entire distribution, but rather a particular parameter of interest, $\psi(P_0)$.

Assuming that the outcome, Y, is a binary event (0=no, 1=yes), the parameter of interest will be one that minimizes a reasonable loss-function. In our case, either by choosing –log(likelihood) loss: $L(O, \psi) = -\log\{\psi(W)^Y(1 - \psi(W))^{1-Y}\}$, or mean squared error-loss $L(O, \psi) = (Y - \psi(W))^2$. For both the minimization of this expected loss under the true P_0 over all functions of W gives: $\psi_0(W) = E_0[Y|W] = argmin_\psi E_0[L(O, \psi)]$, where E_0 refers to the mean under the true data-generating distribution, P_0. This formulation shows that if the objective is to find the best predictor, one needs the best estimate of the risk. In practice, one will find $\hat{\psi} = argmin_{\psi_k, k=1,\dots,K(n)} \hat{E}[L(O, \psi)]$, in other words the best algorithm among a number of candidate ($k = 1, \dots, K(n)$) estimators, where the number $K(n)$ can depend on the sample size. There are ad hoc approaches for this model selection (e.g., AIC, BIC, other fit statistics) and there are objectively optimal ways via cross-validation. The gold standard is to choose an estimate of the risk, $\hat{E}[L(O, \psi)]$, such that when the same procedure is invoked, one obtains the results one would have obtained had we known the true "best" one of the candidates, or the oracle. Among $K(n)$ estimators for the risk, the "oracle" selector is the estimator which minimizes risk under the true

data generating distribution, P_0. Since the true data generating distribution is unknown, this oracle selector is also unknown.

Ideally, one would like a procedure of proposing models and evaluating them that 1) converge to the actual true model as the sample size, n, gets larger, and 2) performs optimally with the sample size at hand. This phrases the problem explicitly and accurately without invoking the finding of the true model as a possibility – just the one closest to the truth among the candidates considered. The basis of such a procedure will be a consistent estimate of the risk.

4.4 Super Learning

Once the parameter of interest has been defined as the minimizer of a risk, estimates are obtained by searching over a model space. Difficulty in searching the entire model space may lead us to pre-specify a collection of candidate algorithms, each searching over a different subspace of the model space. Optimality of the estimators depends on how well each algorithm searches the parameter space. Given a set of algorithms, cross-validation and risk comparison can be used to select the "optimal" estimator. Recent work by van der Laan et al.[60] and Polley and van der Laan[63] has demonstrated that an ensemble of the set of algorithms searching the parameter space is the asymptotically optimal estimator. Van der Laan et al.[60] introduce an algorithm called "super learner" which is a loss based prediction algorithm that combines a collection of prediction algorithms into an ensemble estimator by optimally weighting them using cross-validation. The initial collection of algorithms may differ in various aspects such as the subset of covariates used, the basis functions, the loss functions, etc. Based on established theoretical oracle properties for the cross-validation selector, the super learner performs asymptotically as well as the so-called oracle selector as defined above.[60]

Following Polley and van der Laan[63] we briefly summarize the algorithm in a few steps (readers may refer to van der Laan et al.[60] for a more detailed discussion). For a given set of candidate algorithms: Consider the learning data $X_i = (W_i, Y_i) \sim P_0$, $i = 1, \dots, n$ where Y is the outcome of interest, and W is a p-dimensional set of predictors. The parameter of interest is the conditional probability of CHD given the covariates: $\psi_0 = E[Y|W] = P(Y = 1|W)$. Let $\hat{\psi}_k, k = 1, \dots, K(n)$, be a library of candidate estimators, each a mapping from the empirical probability distribution P_n into the parameter space ψ, where $K(n)$ is the total number of algorithms in this library. The super learner involves the following steps:[63]

1. Fit each algorithm on the entire dataset $X = X_i : i = 1, \dots, n\}$ to estimate $\hat{\psi}_k(W), k = 1, \dots, K(n)$.

2. Split the dataset X into a training and validation sample, according to a V-fold cross-validation scheme.$V = 1, \ldots, v$. Define $T(v)$ to be the v-th training data split and $V(v)$ to be the corresponding validation data split

3. For the v-th fold, fit each algorithm in the library on $T(v)$ and save the predictions on the corresponding validation data, $\hat{\psi}_{k,T(v)}(W_i), X_i \in V(v)$ for $V = 1, \ldots, v$.

4. Stack the predictions from each algorithm together to create a n by K matrix, $Z = \{\hat{\psi}_{k,T(v)}(W_{V(v)}), v = 1, \ldots, V \;\&\; k = 1, \ldots, K\}$

5. Propose a family of weighted combinations of the candidate estimators indexed by weight vector α:

$$m(z|\alpha) = \sum_{k=1}^{K} \alpha_k \hat{\psi}_{k,T(v)}(W_{V(v)}), \alpha_k \geq 0 \; \forall k, \sum_{k=1}^{K} \alpha_k = 1.$$

6. Find the α that minimizes the cross-validated risk of the candidate estimator $\sum_{k=1}^{K} \alpha_k \hat{\psi}_k$:

$$\hat{\alpha} = argmin_{\alpha} \sum_{i=1}^{n} (Y_i - m(z_i|\alpha))^2$$

7. Combine $\hat{\alpha}$ with $\hat{\psi}_k(W), k = 1, \ldots, K(n)$ to create the final super learner fit:

$$\hat{\psi}_{SL}(W) = \sum_{k=1}^{K} \hat{\alpha}_k \hat{\psi}_k(W)$$

Oracle results for the super learner are dependent on the loss function being bounded.[60] The convex combination restriction of the weights implies that if each candidate algorithm is bounded, then the convex combination will be bounded as well. Van der Laan and Dudoit[59] establish oracle results for the cross-validation selector among a set of candidate estimators for general bounded loss functions. Van der Laan et al.[60] apply these results to the super learner. In summary, they establish that if the number of candidate estimators is polynomial in sample size, then the cross- validation selector is either asymptotically equivalent with the oracle selector, or it achieves the parametric rate of convergence log n/n. (see van der Laan et al.[60] and Polley and van der Laan[63])

In what follows, we apply the super learner algorithm to a subset of the Framingham data using various pre-specified algorithms. We demonstrate that semi-parametric estimation of risk of CHD, using the estimation road map introduced earlier and the

super learning algorithm, performs as well as the approaches used for construction of the Framingham models without relying on the investigators' knowledge about the form of the models.

4.5 Application of Super Learning to the Framingham Study for Prediction

A subset of the Framingham Study data was obtained from the National Heart, Lung, and Blood Institute. The data was collected as part of the Framingham study and included information on age, sex, height, weight, blood pressure, diabetes status, smoking status, total cholesterol, HDL-C, and LDL-C for 4,434 participants. The covariates were measured during three examination periods, approximately 6 years apart, roughly from 1956 to 1968. We used the third examination period as the baseline since blood cholesterol was only measured during this examination period. Participants were followed for a total of 24 years (12 years from our baseline) for the outcome of the following events: Angina Pectoris, Myocardial Infarction, Atherothrombotic Infarction or Cerebral Hemorrhage (Stroke) or death.

We categorized hypertension according to JNC-V blood pressure category definitions. TC, HDL-C, and LDL-C were also categorized based on the same cutoffs as in Wilson et al.[58] we defined the outcome of interest as occurrence of any "Hard CHD" (recognized and unrecognized, myocardial infarction, coronary insufficiency, and death due to CHD) in a 10-year follow up period. Any participants who were missing at least one of the risk factors, or dropped out of the study earlier than 10 years were dropped in our analysis, assuming the values were missing at random. After removing the missing values, our data consisted of 1118 men and 1578 women.

To replicate the models derived by Wilson et al.[58], we approached risk modeling two different ways: once we fit a sex-specific logistic regression model with the exact form as the Framingham models, and a second time we applied the super learner to the data. Although the assumption of any difference between sexes should be based on risk comparison as discussed in the road map for estimation, we repeated our analysis for each sex separately to derive sex-specific models similar to Wilson et al.

The following set of algorithms was pre-specified to be used in the super learner: Random Forest[64], K-nearest neighbors[65], neural networks[65], generalized linear models[66], generalized linear model via penalized maximum likelihood[67], generalized additive models[68], and the Deletion/Substitution Algorithm[32].

Table 4.2 Algorithms used in the library of Super Learner for Prediction of CHD

Algorithm	*Description*	*Author*
randomForest	Random Forest	Liaw and Wiener[69]
Nnet	Neural networks	Venables and Ripley[65]
Knn	K-nearest neighbors	Venables and Ripley[65]
glm	generalized linear models	R Development Core Team
Step	stepwise glm with interactions	Hastie & Pregibon[70]
Glmnet	generalized linear models via penalized maximum likelihood	Friedman, Hastie and Tibshirani[67]
gam	generalized additive models	Hastie[68]
DSA	Deletion/Substitution/Addition algorithm	Sinisi and van der Laan[32]

The learning set was divided into 10 different splits for cross validation, each time using $9/10^{th}$ of the data as the training set and the remaining $1/10^{th}$ as the validation set. On each training set, once we estimated the coefficients for the Framingham model, and another time we applied the super learner algorithm. For each validation set, CHD risk scores were predicted and the AUC was calculated based on the models fit to the corresponding training set. We obtained average AUCs for the super learner and the Framingham models by averaging the AUCs obtained from each validation set. It should be noted that we biased the risk estimates downward for the Framingham models since we assumed that model form was pre-specified. For men, the average AUC obtained by the Framingham model was 0.735, and the average AUC obtained for the super learner was equal to 0.737. For women, the average AUC obtained for the Framingham model was equal to 0.728, and the average AUC obtained for super learner was equal to 0.735.

Table 4.3 Super Learner weights calculated for prediction of CHD

Algorithm	*SL Weights for Men*	*SL Weights for Women*
randomForest	0.004	0.106
Nnet	0.000	0.000
Knn	0.000	0.066
glm	0.568	0.000
Step	0.164	0.000
Glmnet, α=0.25	0.116	0.827
Glmnet, α=0.50	0.000	0.000
Glmnet, α=0.75	0.000	0.000
gam	0.145	0.000
DSA	0.000	0.000

4.6 Variable Importance

A secondary goal of the study was to estimate and test the importance of each variable for predicting risk of coronary heart disease. As we described earlier, Wilson et al. approached this goal by testing the coefficients of the sex-specific Cox proportional hazard models. Recently, van der Laan[71], proposed a new approach to variable importance. Based on this approach, one defines the variable importance as a real valued parameter of the data generating distribution (inspired by the causal inference framework of chapters 2 and 3), and uses locally efficient semi-parametric estimators of variable importance that are specifically targeted toward the estimation of this parameter.
In particular, for the Framingham study we defined the variable importance parameter of interest as:

$$P \to \Psi(P)(a) \equiv E_W\big[E[Y|A=a,W]\big]/E_W\big[E[Y|A=0,W]\big] \overset{ass}{=} E[Y_{A=a}]/E[Y_{A=0}]$$

for a particular level of the variable of interest ($A=a$), compared to the baseline group ($A=0$), where the last equality refers to a ratio of so-called counterfactual means. Note that this last step comes from certain assumptions on a causal graph or a non-parametric structural equation model.[17] If the assumptions are true, this parameter can be interpreted as the marginal relative risk of the outcome, had everyone had their A set to $A=a$, adjusting for other variables, W, compared to had everyone had their A set to the baseline value $A=0$, adjusting for other variables W. The parameter is only well defined if both $P(A=a|W)>0$ and $P(A=0|W)>0$.[70] We will refer to violations of this assumption as Experimental Treatment Assignment (ETA) bias.[30,31]

As suggested, this variable importance parameter is inspired by the causal inference framework and under additional assumptions may have a causal interpretation. The reason for choosing this parameter was to compare the estimates to relative risks obtained from a parametric model. As part of our analysis we fit a log-linear model to the data and compared the exponentiated coefficients to our parameter estimates.
For estimation of the above importance variable, one can rely on the super learner as discussed above. However, we used the simpler DSA algorithm, to save computation time. As described in chapters 3 and 4, DSA is a data-adaptive model selection algorithm based on cross-validation. The algorithm selects from a set of candidate generalized linear models that consist of polynomials of the covariates and their tensor products by using three different moves: deletions, substitutions, and additions. DSA was used earlier in the chapter as one of the algorithms in the super learner library for prediction of risk of CHD.

To estimate the parameter of interest, indicator variables were created for each level of the categorical variables. The variable importance measure for each level was estimated by comparing it to a baseline level. The model space was restricted to the models of the form $logit(E[Y|A,W]) = \beta_0 + \beta_1 A + \beta_2 f(A,W)$($\beta_2$ may be a vector). In other words, we modeled occurrence of CHD using a logistic regression model by forcing the exposure variable of interest into the model and allowing the DSA

algorithm to select the rest of the predictors. For categorical variables such as blood pressure, all the indicator groups were forced into the model with the exception of the baseline.
To obtain MLE estimates of the relative risk, we estimated the parameter of interest by the following formula (referred to as G-computation formula[42], see chapters 3,4):

$$\Psi(\hat{P})(a) = \hat{E}_W\left[\hat{E}[Y|A=1,W]\right] / \hat{E}_W\left[\hat{E}[Y|A=0,W]\right] = \frac{1}{n}\sum_{i=1}^{n}\hat{Q}(A=1,W_i) \Big/ \frac{1}{n}\sum_{i=1}^{n}\hat{Q}(A=0,W_i)$$

where $\hat{Q}(A,W) = \hat{E}[Y|A,W]$. We first plugged in 1 for everyone's exposure level of interest, (and 0 for all the other categories), and calculated the marginal mean in the numerator by taking the average predicted probabilities over all participants, and a second time, we plugged in 0 for all the variables (baseline category), and calculated the marginal mean of the predicted probabilities in the denominator. As an example, to estimate the relative risk for having high-normal blood pressure category compared to optimal, we modeled CHD by forcing in all blood pressure categories into the model except the optimal (baseline) category, and allowed the DSA algorithm to pick the rest of the terms using 5-fold cross-validation. We then plugged in 1 for everyone's high-normal category, and 0 for all the other categories, and averaged the predicted probabilities to obtain an estimate of the marginal mean in the numerator. The second time, we plugged in 0 into all the blood pressure categories, and averaged the predicted probabilities to obtain the marginal mean in the denominator.

In the next step of our analysis, we used TMLE to estimate the relative risks by targeting the marginal means in the numerator and the denominator separately. As described in chapters 3, and 4, TMLE modifies MLE by adding a clever covariate to the model, treating the initial MLE estimator as an offset, in a way that reduces bias for the target parameter of interest. We noted that, a plug-in estimator for the parameter of the density estimator may be biased due to model misspecification (unless the estimate is nonparametric). In such cases, the TMLE directly addresses the bias issue by carrying out a subsequent clever parametric maximum likelihood fit that is directly tailored to remove bias for the target parameter of interest, treating the initial MLE estimator as an offset. In addition to being asymptotically efficient, the resulting estimator will also be double robust.[25,33,34] The TMLE estimator for the marginal relative risk is given by:

$$\Psi(\hat{P})(a) = \frac{1}{n}\sum_{i=1}^{n}\hat{Q}^*(A=1,W_i) \Big/ \frac{1}{n}\sum_{i=1}^{n}\hat{Q}^*(A=0,W_i)$$

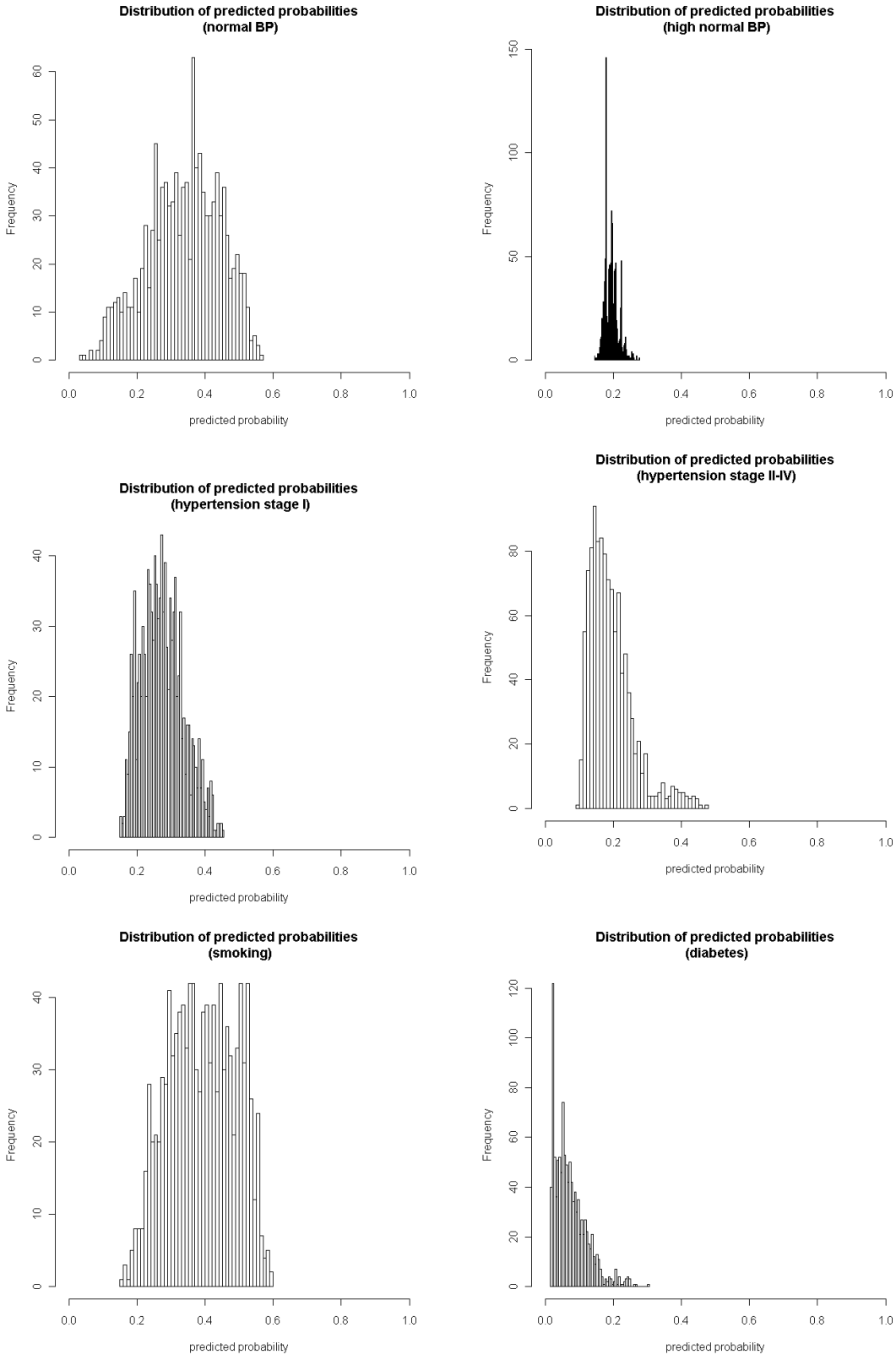
where $\hat{Q}^*(A,W_i) = logit^{-1}(\hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 f(A,W) + \hat{\varepsilon} h(A,W))$ is the updated model, $\hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 f(A,W)$ ($\hat{\beta}_2$ may be a vector of parameters) is the MLE model, and $h(A,W) = \frac{I(A=a)}{P(A=a|W)}$ is the clever covariate. The probabilities in the denominator of the clever covariate, $P(A=a|W)$, can be estimated either by assuming a model and

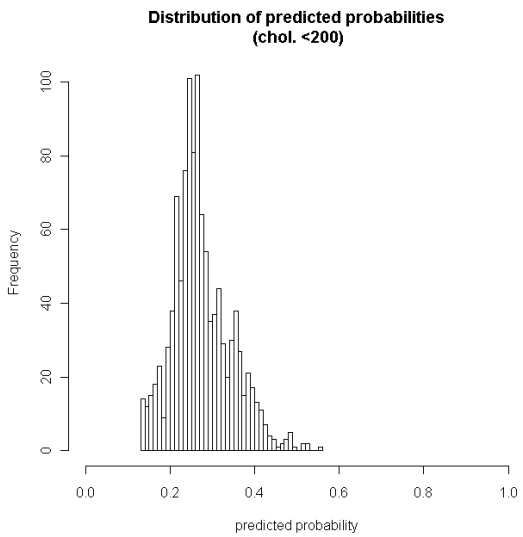
estimating the parameter (e.g. logistic regression), or by performing model selection (see van der Laan and Gruber[36]).

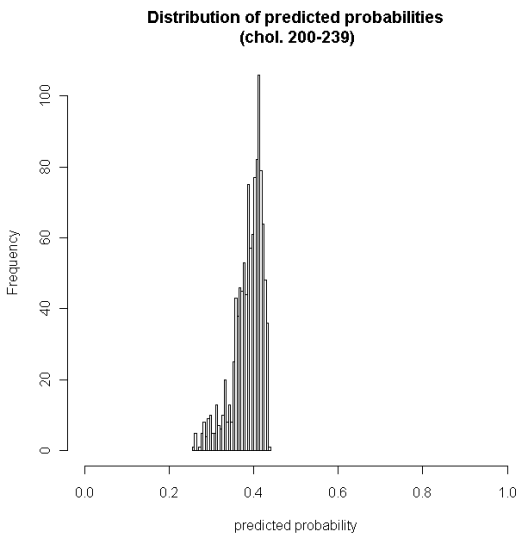
To obtain the TMLE estimates, we modeled the exposures using a generalized linear models with all the other variables fixed into the model. If the exposure was binary (e.g. diabetes) we fit a logistic regression model, and if the exposure was categorical we fit a multinomial logistic regression (before creating indicators) to predict the probabilities for exposure categories. For each exposure variable we formed the clever covariate, $\frac{I(A=a)}{P(A=a|W)}$, and updated the corresponding model for CHD by adding the clever covariate in the model and treating the initial MLE estimator as an offset.

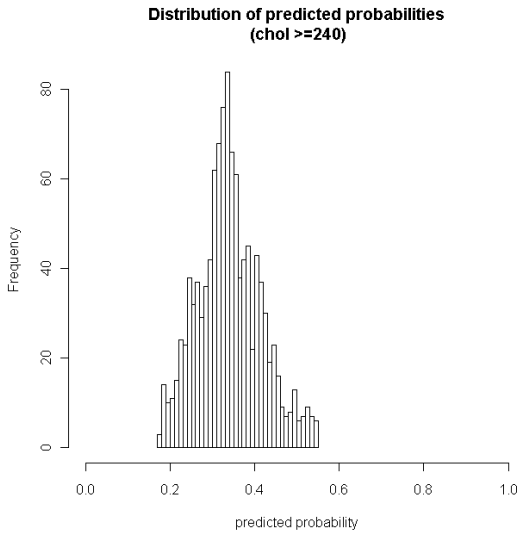
We repeated the steps in estimation of the marginal means as in MLE estimation with the difference that the updated CHD model was used to obtain the predicted probabilities. Note that small predicted probabilities in the denominator of the clever covariates may result in large values for the clever covariate which in turn results in inflation of the variance. This may occur due to the ETA assumption being violated. To check for such violations we plotted the predicted probabilities from the exposure models. The probabilities need to be bounded away from 0 and 1. (see figure 4.1)

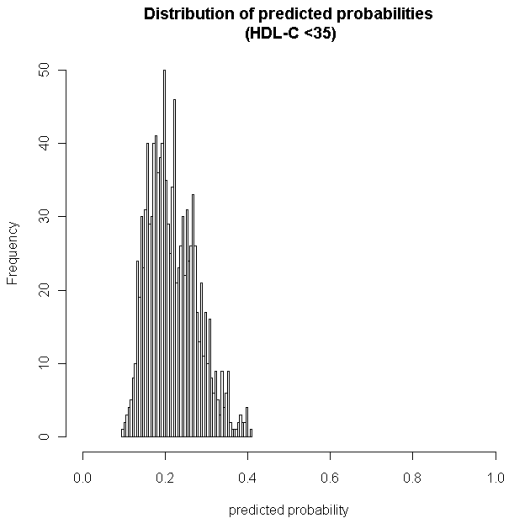
Figure 4.1 Distribution of the predicted probabilities for the risk factors in the Framingham study

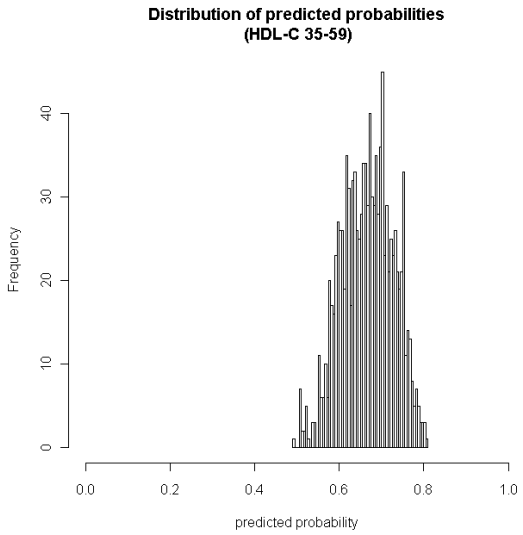


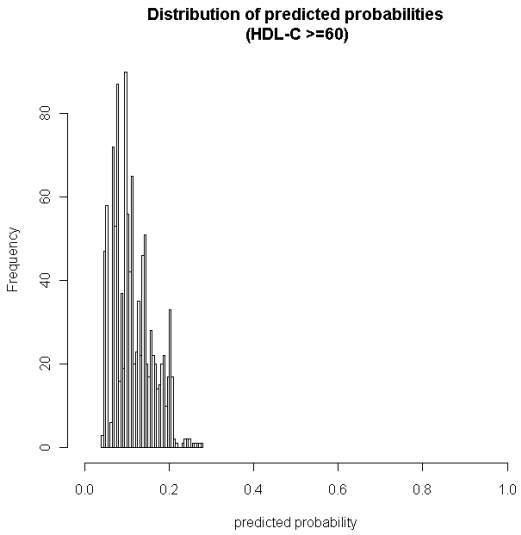

Distribution of predicted probabilities
(chol. <200)
Frequency
predicted probability


Distribution of predicted probabilities
(chol. 200-239)
Frequency
predicted probability


Distribution of predicted probabilities
(chol >=240)
Frequency
predicted probability


Distribution of predicted probabilities
(HDL-C <35)
Frequency
predicted probability


Distribution of predicted probabilities
(HDL-C 35-59)
Frequency
predicted probability


Distribution of predicted probabilities
(HDL-C >=60)
Frequency
predicted probability

To compare our estimates to similar ones obtained from a parametric model, we fit a log-linear model with all the variables used in the Framingham study included in the model. The exponentiated coefficients from this log-linear model provide adjusted relative risk measures of importance comparable to the ones defined earlier. To show this, we use a log-linear model and the formula we used for MLE as follows:

$$log\left(\hat{E}[Y|A=1,W]\right) = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 W,$$
$$log\left(\hat{E}[Y|A=0,W]\right) = \hat{\beta}_0 + \hat{\beta}_2 W.$$

Thus,

$$\hat{E}[Y|A=1,W_i] = e^{(\hat{\beta}_0+\hat{\beta}_1)} e^{(\hat{\beta}_2 W_i)},$$
$$\hat{E}[Y|A=0,W_i] = e^{\hat{\beta}_0} e^{(\hat{\beta}_2 W_i)}.$$

So,

$$\frac{1}{n}\sum_{i=1}^{n} \hat{E}[Y|A=1,W_i] = e^{(\hat{\beta}_0+\hat{\beta}_1)} \frac{1}{n}\sum_{i=1}^{n} e^{(\hat{\beta}_2 W_i)},$$
$$\frac{1}{n}\sum_{i=1}^{n} \hat{E}[Y|A=0,W_i] = e^{\hat{\beta}_0} \frac{1}{n}\sum_{i=1}^{n} e^{(\hat{\beta}_2 W_i)}.$$

It follows that

$$\hat{E}_W\left[\hat{E}[Y|A=1,W]\right] \Big/ \hat{E}_W\left[\hat{E}[Y|A=0,W]\right] = \frac{e^{(\hat{\beta}_0+\hat{\beta}_1)} \frac{1}{n}\sum_{i=1}^{n} e^{(\hat{\beta}_2 W_i)}}{e^{\hat{\beta}_0} \times \frac{1}{n}\sum_{i=1}^{n} e^{(\hat{\beta}_2 W_i)}} = e^{\hat{\beta}_1}$$

Finally, we used non-parametric bootstrap to obtain robust standard errors. We re-sampled our data 500 times. For each bootstrap sample we obtained MLE, TMLE, and parametric estimates as described in this section.

4.7 Results of Variable Importance Analysis

Table 4.4 contains the bootstrap estimates obtained from fitting a log-linear model. Based on this table, relative risks for having stage I or higher hypertension (compared to normal blood pressure) and smoking were statistically significantly different from 1 in men. For women, the relative risks for having stage II or higher hypertension, smoking, and diabetes were statistically significantly different from 1.

Table 4.4 Adjusted parametric estimates based on 500 bootstrap samples

		Men		Women	
		Relative Risk	95% CI	Relative Risk	95% CI
Blood Pressure	Normal	1		1	
	High Normal	1.292(0.216)	(0.960, 1.640)	1.023(0.243)	(0.644, 1.444)
	Hyper I	1.811(0.255)	(1.406, 2.229)	1.379(0.243)	(0.993, 1.812)
	Hyper II-IV	1.948(0.293)	(1.524, 2.469)	1.775(0.317)	(1.321, 2.333)
Smoke	Yes/No	1.505(0.153)	(1.254, 1.762)	1.719(0.239)	(1.325, 2.121)
Diabetes	Yes/No	1.145(0.167)	(0.884, 1.428)	2.233(0.339)	(1.723, 2.802)
TC (mg/dL)					
	<200	1		1	
	200-239	1.082(0.129)	(0.885, 1.301)	0.873(0.166)	(0.635, 1.176)
	>=240	1.021(0.121)	(0.830, 1.229)	0.885(0.157)	(0.660, 1.160)
HDL-C (mg/dL)					
	<35	1.125(0.133)	(0.907, 1.352)	1.220(0.220)	(0.885, 1.626)
	35-59	1		1	
	>=60	1.025(0.157)	(0.757, 1.277)	0.950(0.134)	(0.725, 1.174)

Table 4.5 illustrates the MLE estimates obtained from the bootstrap samples. Similar to the parametric estimates, the relative risks for having stage I or higher hypertension and smoking were statistically significantly different from 1. Furthermore, the relative risk for having the highest level of cholesterol (>=240 mg/dL) compared to the baseline(<200 mg/dL) was statistically significantly different from 1 for men. This change in significance was mainly due to a rise in the magnitude of the relative risk when estimated using MLE (as there is also an increase in the standard error).

Table 4.5 Semi-parametric Maximum Likelihood Estimates based on 500 bootstrap samples

		Men		Women	
		Relative Risk	95% CI	Relative Risk	95% CI
Blood Pressure	Normal	1		1	
	High Normal	1.268(0.202)	(0.987,1.636)	0.996(0.219)	(0.668, 1.379)
	Hyper I	1.792(0.260)	(1.427,2.263)	1.332(0.237)	(0.996, 1.751)
	Hyper II-IV	1.930(0.285)	(1.506,2.432)	1.755(0.304)	(1.306, 2.294)
Smoke	Yes/No	1.453(0.143)	(1.219,1.700)	1.628(0.213)	(1.310, 2.012)
Diabetes	Yes/No	1.173(0.190)	(0.863,1.517)	2.476(0.404)	(1.880, 3.156)
TC (mg/dL)					
	<200	1		1	
	200-239	0.993(0.145)	(0.740,1.248)	0.808(0.207)	(0.526,1.144)
	>=240	1.355(0.160)	(1.105,1.630)	1.063(0.232)	(0.746, 1.451)
HDL-C (mg/dL)					
	<35	1.083(0.128)	(0.879,1.304)	1.226(0.245)	(0.846,1.658)
	35-59	1		1	
	>=60	1.025(0.163)	(0.768,1.301)	0.977(0.140)	(0.764, 1.208)

Finally, the results obtained from TMLE are shown in table 4.6. Similar to the parametric results and the MLE results, the relative risk for having stage I or higher hypertension and smoking were statistically different from 1 in men. Additionally, the relative risk for being diabetic was significantly different from 1. This was due to an increase in the magnitude of the relative risk when estimated using TMLE. Similar to the parametric results, and unlike the MLE estimates, the relative risk for having high TC was not statistically different from 1 in men.

For women we observed similar results to the parametric and MLE estimates. The relative risk for having stage II or higher hypertension, smoking, and diabetes were statistically significantly different from 1. However, the magnitude of relative risk estimated for diabetes in women using TMLE (RR=1.47) was smaller than both the MLE estimate (RR=2.47), and the parametric estimate (RR=2.23).

Table 4.6 Semi-parametric Targeted Maximum Likelihood Estimates based on 500 bootstrap samples

		Men		Women	
		Relative Risk	95% CI	Relative Risk	95% CI
Blood Pressure	Normal	1		1	
	High Normal	1.269(0.207)	(0.994,1.653)	0.987(0.221)	(0.653,1.387)
	Hyper I	1.794(0.262)	(1.418,2.260)	1.331(0.239)	(0.992,1.744)
	Hyper II-IV	1.933(0.286)	(1.503,2.438)	1.753(0.305)	(1.305,2.282)
Smoke	Yes/No	1.460(0.145)	(1.237,1.702)	1.609(0.210)	(1.301,1.996)
Diabetes	Yes/No	1.386(0.206)	(1.060,1.690)	1.477(0.333)	(1.055,2.013)
TC (mg/dL)					
	<200	1		1	
	200-239	1.076(0.153)	(0.851,1.334)	0.818(0.179)	(0.572, 1.130)
	>=240	0.965(0.193)	(0.611,1.274)	0.852(0.234)	(0.559,1.268)
HDL-C (mg/dL)					
	<35	1.061(0.139)	(0.841,1.320)	1.086(0.283)	(0.632,1.572)
	35-59	1		1	
	>=60	1.024(0.164)	(0.766,1.299)	0.978(0.142)	(0.758,1.210)

4.8 Discussion

Overall, the AUCs obtained from our risk prediction models are similar to those obtained by Wilson et al. The AUC estimated for the super learner for men is exactly the same as the Framingham study (0.73). The AUC for women is slightly lower than the original Framingham study (0.73 compared to 0.76). This difference may be due to over-fitting in the original Framingham study as it appears that Wilson et al. created and evaluated the Framingham models on the same dataset.

Based on our variable importance analysis (using TMLE), we found that the relative risk for having stage I or higher hypertension and smoking were statistically different from 1 in men. In addition, we observed a statistically significant relative risk in men similar to women. This finding corrects the inconsistency we observed in our parametric and the MLE estimates which show a significant relative risk for women but not for men. Wilson et al. had reported relative risks significantly higher than 1 for diabetes in both men and women.

In our analysis, variable importance was defined based on causal inference definitions of marginal relative risk in our analysis. This definition may not necessarily be the most interesting to the investigator. In particular, one may define importance of a variables for prediction purposes as any improvement in the risk estimate, when the variable is included in the model (or the covariate space the machine learning algorithms are searching) versus when the variable is excluded. Significance of this parameter may be formally tested by comparing risk estimates from including and excluding the variable.

In this chapter, we examined alternatives to traditional approaches for statistical modeling using the Framingham study. Parametric models (and in some cases semi-parametric models, i.e. Cox proportional hazard) are nearly always mis-specified, and add unwarranted assumptions to the estimation and interpretation of the parameter of interest. Relying on statistical theory for loss-based estimation, semi-parametric modeling, and robust sampling based inference, allows much flexibility in estimation and interpretation of the parameter of interest. This point was demonstrated in our paper by applying the super learner to the Framingham study for risk prediction. Given the power of computers and availability of machine learning software, loss-based semi-parametric estimation methods will perhaps replace the common use of parametric modeling using ad hoc methods.

References

1. Skrabanek P. The Emptiness of the Black Box. *Epidemiology* 1994; 5:553-555.

2. Greenland S, Gago-Dominguez M, Castelao JE. The Value of Risk-Factor ("Black-Box") Epidemiology. *Epidemiology* 2004; 15: 529–535.

3. Savitz DA. In Defense of Black Box Epidemiology. *Epidemiology* 1994; 5:550-552.
4. Weed DL. Beyond Black Box Epidemiology. *American Journal of Public Health* 1998; 88:12-14.

5. Peto R. The need for ignorance in cancer research. The Encyclopedia of Medical Ignorance Oxford, England, 1984:129-133.

6. Gail M, Brinton LA, Byar DP, et al. Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually. *Journal of the National Cancer Institute* 1989; 81(24): 1879-1886.

7. Mayo DG, Spanos A. When can risk-factor epidemiology provide reliable tests?! [commentary] *Epidemiology*. 2004; 15:523–524.

8. Greenland S, Gago-Dominguez M, Castelao JE. Authors' Response [commentary]. Epidemiology 2004; 15(5): 527–528.

9. Petersen ML, Sinisi SE, Van der Laan MJ. Estimation of direct causal effects. *Epidemiology* 2006; 17: 276 – 84.

10. Pirart J. Diabetes mellitus and its degenerative complications: a prospective study of 4400 patients observed between 1947 and 1973. *Diabetes Care* 1978; 1: 168-188, 252-263.

11. Backonja M, Beydoun A, Edwards KR, et al. Gabapentin for the Symptomatic Treatment of Painful Neuropathy in Patients with Diabetes Mellitus. *Journal of American Medical Association* 1998; 280: 1831-1836.

12. Beecher HK. The powerful placebo. *Journal of American Medical Association* 1955; 159:1602-1606.

13. Hrobjartsson A, Gotzsche PC. Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment. *New England Journal of Medicine* 2001; 344:1594-602.

14. Turner JA, Deyo RA, Loeser JD, et al. The importance of placebo effects in pain treatment and research. *Journal of American Medical Association* 1994; 271:1609-14

15. Miller FG, Rosenstein LD. Variance and dissent: The nature and power of the placebo effect. *Journal of Clinical Epidemiology* 2006; 59:331-335.

16. Petkova E, Tarpey T, Govindarajulu U. Predicting Potential Placebo Effect in Drug Treated Subjects. *The International Journal of Biostatistics* 2009; 5(1): 23. Available from: http://www.bepress.com/ijb/

17. Pearl, J. Causality: models, reasoning, and inference. Cambridge University Press, New York, NY, 2000.

18. Lynch KG, Cary M, Gallop R, Ten Have TR. Causal mediation analyses for randomized trials. *Health Services and Outcomes Research Methodology* 2008; 8: 57-76.

19. Rosenblum M, Jewell NP, van der Laan M, Shiboski S, van der Straten A, Padian N. Analysing direct effects in randomized trials with secondary interventions: an application to human immunodeficiency virus preventio trials. *Journal of the Royal Statistical Society A* 2009; 172: 443-465.

20. Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* 2000; 11(5):561–570.

21. Robins JM. Statistical models in epidemiology, the environment, and clinical trials: Marginal structural models versus structural nested models as tools for causal inference. New York: Springer; 2000b. p. 95–133.

22. Robins JM, Rotnitzky A. Comment on the Bickel and Kwon article, "Inference for semiparametric models: Some questions and an answer". *Statistica Sinica* 2001; 11(4):920–936.

23. Robins JM, Rotnitzky A, Van der Laan MJ. Comment on "On Profile Likelihood" by S.A. Murphy and A.W. van der Vaart. *Journal of the American Statistical Association – Theory and Methods* 2000; 450:431–435.

24. Robins JM. Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association* 2000a.

25. Van der Laan MJ, Rubin D. Targeted maximum likelihood learning. *The International Journal of Biostatistics* 2006; 2(1). Available from: http://www.bepress.com/ijb/

26. Scharfstein DO, Rotnitzky A, and Robins JM. Adjusting for non-ignorable drop-out using semiparametric nonresponse models, (with Discussion and Rejoinder). *Journal of the American Statistical Association* 1999; 94:1096-1120, 1121-1146.

27. Jewell, NP. *Statistics for Epidemiology*. Chapman & Hall/CRC Press, Boca Raton, 2003.

28. Van der Laan MJ, Robins JM. Unified Methods for Censored Longitudinal Data and Causality. Springer-Verlag, New York, NY, 2003.

29. Rubin DB. Bayesian inference for causal effects*: the role of randomization. Annals of Statistics 1978; 7:34–58.*

30. Neugebauer R, Van der Laan MJ. Why Prefer Double Robust Estimates? Illustration with Causal Point Treatment Studies. 2002; Available from: http://works.bepress.com/mark_van_der_laan/181/

31. Dudoit S, Van der Laan MJ. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology* 2005; 2(2):131-154.

32. Sinisi S, Van der Laan MJ. The deletion/substitution/addition algorithm in loss function based estimation: Applications in genomics. *Journal of Statistical Methods in Molecular Biology* 2004; 3(1)

33. Van der Laan MJ, Gruber S. Collaborative double robust penalized targeted maximum likelihood estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series* 2009; 246. Available from: http://www.bepress.com/ucbbiostat/paper246

34. Gruber S, van der Laan MJ. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *U.C. Berkeley Division of Biostatistics Working Paper Series* 2009; 265. Available from: http://www.bepress.com/ucbbiostat/paper265

35. Turner JA, Jensen MP, Warms CA, Cardenas DD Blinding effectiveness and association of pretreatment expectations with pain improvement in a double-blind randomized controlled trial. *Pain* 2002; 99: 91-99.

36. Szmuness W, Stevens CE, Harley EJ, et al. Hepatitis B vaccine: demonstration of efficacy in a controlled clinical trial in a high-risk population in the United States. *New England Journal of Medicine* 1980; 303:833–841.

37. Bartholow BN, Buchbinder S, Celum C, et al. HIV sexual risk behavior over 36 months of follow-up in the world's first HIV vaccine efficacy trial. *Journal of Acquired Immune Deficiency Syndrome* 2005; 39: 90–101.

38. Van der Laan MJ. Targeted Maximum Likelihood Based Causal Inference: Part I. *The International Journal of Biostatistics* 2010; 6(2). Available from: http://www.bepress.com/ijb/

39. Van der Laan MJ. Targeted Maximum Likelihood Based Causal Inference: Part II. *The International Journal of Biostatistics* 2010; 6(2). Available from: http://www.bepress.com/ijb/

40. Neyman J (1923). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9, translated in *Statistical Science*, (with discussion) 1990; 5(4): 465-480

41. Rubin DB. Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies. *Journal of Educational Psychology* 1974; 66: 688-701.

42. Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect. *Mathematical Modeling* 1986; 7:1393–1512.

43. Holland P. Statistics and Causal Inference. *Journal of the American Statistical Association* 1986; 81: 945-961.

44. Robins J, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 1992; 3:143-155.

45. Robins JM. Semantics of causal dag models and the identification of direct and indirect effects. In N. Hjort P. Green and S. Richardson, editors, *Highly Structured Stochastic Systems*. Oxford University Press, Oxford, 2003.

46. Petersen ML, Sinisi SE, van der Laan MJ. Estimation of direct causal effects. *Epidemiology* 2006;17:276-84.

47. Bickel PJ, Klaassen CAJ , Ritov Y,et al.Efficient and Adaptive Estimation for Semiparametric Models. 1993 Johns Hopkins University Press,Baltimore.

48. Tsiatis AA. Semiparametric theory and missing data. Springer, 2006.

49. Ney PG, Collins C,Spensor C.Double blind: double talk or are there ways to do better research. Medical Hypothesis 1986; 21:119–126.

50. Bang H, Flaherty SP, Kolahi J, et al. Blinding assessment in clinical trials: A review of statistical methods and a proposal of blinding assessment protocol. *Clinical Research and Regulatory Affairs* 2010; 27(2):42-51.

51. Colford JM, Rees JR, Wade TJ, et al. Participant blinding and gastrointestinal illness in a randomized, controlled trial of an in-home drinking water intervention. *Emerging Infectious Diseases* 2002; 8 : 29–36.

52. James KE, Bloch DA, Lee KK,et al. An index for assessing blindness in a multi-centre clinical trial: disulfiram for alcohol cessation--a VA cooperative study. *Statistics in Medicine* 1996; 15(13):1421–1434.

53. Howard J, Whittemore AS, Hoover JJ, et al. How blind was the patient blind in AMIS. *Clinical Pharmacology and Therapeutics,* 1982;32:543–553.

54. Bang H, Liyun N, Davis C. Assessment of blinding in clinical trials. *Controlled Clinical Trials* 2004;25:143–156.

55. Chambaz A, Pierre N, van der Laan MJ. Estimation of a Non-Parametric Variable Importance Measure of a Continuous Exposure. *U.C. Berkeley Division of Biostatistics Working Paper* Series 2011. Paper 292. Available at http://www.bepress.com/ucbbiostat/paper292

56. Van der Laan MJ, Hubbard AE, Jewell NP. Learning from data: Semiparametric models versus faith-based inference. *Epidemiology* 2010; 21:479–481.

57. Keenan NL, Shaw KM. Coronary heart disease and stroke deaths—United States,2006. *Morbidity Mortality Weekly Report* 2011; 60(1): 62–66.

58. Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;97:1837-47.

59. Van der Laan MJ, Dudoit S. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. 2003. Available from:

URL http://www.bepress.com/ucbbiostat/paper130/.

60. Van der Laan, Polley EC, Hubbard AE. Super learner. *Statistical Applications in Genetics and Molecular Biology*.2007; 6(25): Article 25

61. Van der Laan MJ, Rose S. Targeted Learning: Causal Inference for Observational and Experimental Data. Springer, 2011

62. Gail MH, Pfeiffer RM. On criteria for evaluating models of absolute risk. *Biostatistics* 2005; 6:227-239.

63. Polley EC, van der Laan MJ. Super Learner In Prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series* 2010; Paper 266.Available at http://www.bepress.com/ucbbiostat/paper266.

64. Breiman L. Random Forests. Machine Learning 2001; 45(1): 5–32.

65. Venables WN, Ripley BD. Modern Applied Statistics with S. Springer, New York, fourth
edition, 2002
66. Nelder J, Wedderburn R. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A* 1972;135(3):370–384.
67. Friedman J, Hastie T, Tibshirani R. glmnet: Lasso and elastic-net regularized generalizedlinear models, 2010. Available from http://CRAN.R-project.org/package=glmnet.

68. Hastie TJ. Generalized additive models. In J. M. Chambers and T. Hastie, editors, Statistical Models in S, chapter 7. Wadsworth & Brooks/Cole, 1992.

69. Liaw A Wiener M. Classification and regression by randomforest. R News 2002;2(3):18–22. Available at http://CRAN.R-project.org/package=randomForest

70. Hastie TJ, Pregibon D. Generalized linear models. Chapter 6 of Statistical Models in S eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole 1992

71. Van der Laan MJ. Statistical inference for variable importance. International *Journal of Biostatistics* 2006; 2(1).