

# Preface

The statistics profession is at a unique point in history. The need for valid statistical tools is greater than ever; data sets are massive, often measuring hundreds of thousands of measurements for a single subject. The field is ready for a revolution, one driven by clear, objective benchmarks under which tools can be evaluated.

Statisticians must be ready to take on this challenge. They have to be dynamic and thoroughly trained in statistical concepts. More than ever, statisticians need to work effectively in interdisciplinary teams and understand the immense importance of objective benchmarks to evaluate statistical tools developed to learn from data. They have to produce energetic leaders who stick to a thorough a priori road map, and who also break with current practice when necessary.

Why do we need a revolution? Can we not keep doing what we have been doing? Sadly, nearly all data analyses are based on the application of so-called parametric (or other restrictive) statistical models that assume the data-generating distributions have specific forms. Many agree that these statistical models are wrong. That is, everybody knows that linear or logistic regression in parametric statistical models and Cox proportional hazards models are specified incorrectly. In the early 1900s, when R.A. Fisher developed maximum likelihood estimation, these parametric statistical models were suitable since the data structures were very low dimensional. Therefore, saturated parametric statistical models could be applied. However, today statisticians still use these models to draw conclusions in high-dimensional data and then hope these conclusions are not too wrong.

It is too easy to state that using methods we know are wrong is an acceptable practice: it is not!

The original purpose of a statistical model is to develop a set of realistic assumptions about the probability distribution generating the data (i.e., incorporating background knowledge). However, in practice, restrictive parametric statistical models are essentially always used because standard software is available. These statistical models also allow the user to obtain  $p$ -values and confidence intervals for the tar-

get parameter of the probability distribution, which are desired to make sense out of data. Unfortunately, these measures of uncertainty are even more susceptible to bias than the effect estimator. We know that for large enough sample sizes, every study, including one in which the null hypothesis of no effect is true, will declare a statistically significant effect.

Some practitioners will tell you that they have extensive training, that they are experts in applying these tools and should be allowed to choose the statistical models to use in response to the data. Be alarmed! It is no accident that the chess computer beats the world chess champion. Humans are not as good at learning from data and are easily susceptible to beliefs about the data.

For example, an investigator may be convinced that the probability of having a heart attack has a particular functional form – a function of the dose of the studied drug and characteristics of the sampled subject. However, if you bring in another expert, his or her belief about the functional form may differ. Or, many statistical model fits may be considered, dropping variables that are nonsignificant, resulting in a particular selection of a statistical model fit. Ignoring this selection process, which is common, leaves us with faulty inference.

With high-dimensional data, not only is the correct specification of the parametric statistical model an impossible challenge, but the complexity of the parametric statistical model also may increase to the point that there are more unknown parameters than observations. The true functional form also might be described by a complex function not easily approximated by main terms.

For these reasons, allowing humans to include only their true, realistic knowledge (e.g., treatment is randomized, such as in a randomized controlled trial, and our data set represents  $n$  independent and identically distributed observations of a random variable) is essential. That is, instead of assuming misspecified parametric or heavily restrictive semiparametric statistical models, and viewing the (regression) coefficients in these statistical models as the target parameters of interest, we need to define the statistical estimation problem in terms of nonparametric or semiparametric statistical models that represent realistic knowledge, and in addition we must define the target parameter as a particular function of the true probability distribution of the data. This changes the game in a dramatic way relative to current practice; one starts thinking about real knowledge in terms of the underlying experiment that generated the data set and what the real questions of interest are in terms of a feature of the data-generating probability distribution.

The concept of a statistical model is very important, but we need to go back to its true meaning. We need to be able to incorporate true knowledge in an effective way. In addition, we need to develop and use data-adaptive tools for all parameters of the data-generating distribution, including parameters targeting causal effects of interventions on the system underlying the data-generating experiment. The latter typically represent our real interest. We are not only trying to sensibly observe, but also to learn how the world operates.

What about machine learning, which is concerned with the development of black-box algorithms that map data (and few assumptions) into a desired object? For example, an important topic in machine learning is prediction. Here the goal is to

map the data, consisting of multiple records with a list of input variables and an output variable, into a prediction function that can be used to map a new set of input variables into a predicted outcome. Indeed, this is in sharp contrast to using misspecified parametric statistical models. However, the goal is often a whole prediction function, and the machines are tailored to fit this whole prediction function. As a consequence, these methods are too biased (and not grounded by efficiency theory) for a particular effect of interest. Typical complexities in the data such as missingness or censoring have also received little attention in machine learning. In addition, statistical inference in terms of assessment of uncertainty (e.g., confidence intervals) is typically lacking in this field.

Even in machine learning there is often unsupported devotion to beliefs, in this case, to the belief that certain algorithms are superior. No single algorithm (e.g., random forests, support vector machines, etc.) will always outperform all others in all data types, or even within specific data types (e.g., SNP data from genomewide association studies). One cannot know a priori which algorithm to choose. It's like picking the student who gets the top grade in a course on the first day of class.

The tools we develop must be grounded in theory, such as an optimality theory, that shows certain methods are more optimal than others and, in addition, should be evaluated with objective benchmark simulation studies. For example, one can compare methods based on mean squared error with respect to the truth. It is not enough to have tools that use the data to fit the truth well. We also require an assessment of uncertainty (e.g., confidence intervals), the very backbone of statistical learning. That is, we cannot give up on the reliable assessment of uncertainty in our estimates.

Examples of new methodological directions in statistical learning satisfying these requirements include (1) the full generalization and utilization of cross-validation as an estimator selection tool so that the subjective choices made by humans are now made by the machine and (2) targeting the fitting of the probability distribution of the data toward the target parameter so that the mean squared error of the substitution estimator of the target parameter with respect to the target parameter is optimized. Important and exciting statistical research areas where new developments are taking place in response to the nonvalidity of the previous generation of tools are: adaptive designs in clinical trials and observational studies, multiple and group sequential testing, causal inference, and Bayesian learning in realistic semiparametric statistical models, among others.

Statisticians cannot be afraid to go against standard practice. Remaining open to, interested in, and a developer of newer, sounder methodology is perhaps the one key thing each statistician can do. We must all continue learning, questioning, and adapting as new statistical challenges are presented.

The science of learning from data (i.e., statistics) is arguably the most beautiful and inspiring field – one in which we try to understand the very essence of human beings. However, we should stop fooling ourselves and actually design and develop powerful machines and statistical tools that can carry out specific learning tasks. There is no better time to make a truly meaningful difference.<sup>1</sup>

---

<sup>1</sup> A version of this content originally appeared in the September 2010 issue of *Amstat News*, the membership magazine of the American Statistical Association (van der Laan and Rose 2010).

## The Journey

**Mark:** I view targeted maximum likelihood estimation (TMLE), presented in this book, as the result of a long journey, starting with my Ph.D. research up to now. We hope that the following succinct summary of this path towards a general toolbox for statistical learning from data will provide the reader with useful perspective and understanding.

During my Ph.D. work (1990–1993) under the guidance of Dr. Richard Gill, I worked on the theoretical understanding of the maximum likelihood estimator for semiparametric statistical models, with a focus on the nonparametric maximum likelihood estimator of the bivariate survival distribution function for bivariate right-censored survival times and a nonparametric statistical model for the data-generating distribution. This challenging bivariate survival function estimation problem demonstrated that the nonparametric maximum likelihood estimator easily fails to be uniquely defined, or fails to approximate the true data-generating distribution for large sample sizes. That is, for realistic statistical models for the data-generating distribution, and even for relatively low-dimensional data structures, the maximum likelihood estimator is often ill defined and inconsistent for target parameters, and regularization through smoothing or stratification is necessary to repair it. It also demonstrated that, for larger dimensional data structures, the repair of maximum likelihood estimation in nonparametric statistical models through smoothing comes at an unacceptable price with respect to finite sample performance.

Right after completing my Ph.D., I met Dr. James M. Robins, whose research focused on estimation with censored data and, in particular, estimation of causal effects of time-dependent treatment regimens on an outcome of interest based on observing replicates of high-dimensional longitudinal data structures in the presence of informative missingness and dropout and time-dependent confounding of the treatment. This was an immensely exciting time, and a whole new world opened up for me. Instead of working on toy extractions of real-world problems, Robins and his colleagues worked on solving the actual estimation problems as they occur in practice, avoiding convenient simplifications or assumptions. The work of Robins' group made clear that statistical learning was far beyond the world of standard software and the corresponding practice of statistics based on restrictive parametric statistical models, and also far beyond the world of maximum likelihood estimation for semiparametric statistical models.

Concepts such as coarsening at random, orthogonal complement of the nuisance tangent space of a target parameter, estimating functions for the target parameter implied by the latter, double robustness of these estimating functions and their corresponding estimators, locally efficient estimators of the target parameter, and so on, became part of my language. As a crown on our collaborations, in 2003 we wrote a book called *Unified Methods for Censored Longitudinal Data and Causality*. This book provided a comprehensive treatment of the estimating equation methodology for estimation of target parameters of the data-generating distribution in semiparametric statistical models, demonstrated on complex censored and longitudinal (causal inference) data structures.

From a person trying to repair maximum likelihood estimation, I had become a proponent for estimating equation methodology, a methodology that targets the parameter of interest instead of the maximum likelihood estimation methodology, which aims to estimate the whole distribution of the data. When writing the book in 2003, some nonnatural hurdles occurred and we proposed no solutions for them. To start with, the optimal estimating function for the target parameter might not exist since the efficient influence curve, though a function of the distribution of the data on the unit, cannot necessarily be represented as an (estimating) function in the target parameter of interest and a variation-independent nuisance parameter. If we ignored this first hurdle, we were still left with the following hurdles. Estimators defined by a solution of an estimating equation (1) might not exist, (2) might be nonunique due to the existence of multiple solutions, (3) are not substitution estimators and thus do not respect known statistical model constraints, and (4) are sensitive to how the nuisance parameter (that the estimating function depends on) is estimated, while a good fit of the nuisance parameter itself is not a good measure for its role in the mean squared error of the estimator of the target parameter.

These hurdles, which also affect the practical performance and robustness of the estimators, made it impossible to push this impressive estimating equation methodology forward as a general statistical tool to replace current practice. It made me move back towards substitution estimators using methods based on maximizing or minimizing an empirical criterion such as the maximum likelihood estimator, and plugging in the resulting estimator in the target parameter mapping that maps a probability distribution of the data into the desired target parameter.

Specifically, additional research we conducted in 2003 proposed a unified loss-based learning methodology (van der Laan and Dudoit 2003). The methodology was based on defining a (typically infinite-dimensional) parameter of the probability distribution of the data as a minimizer of the expectation of a loss function (e.g., log-likelihood or squared error loss function) and the aggressive utilization of cross-validation as a tool to select among candidate estimators. The loss function for the desired part of the probability distribution of the data was also allowed to be indexed by an unknown nuisance parameter, thereby making this methodology very general, including prediction or density estimation based on general censored data structures.

The general theoretical optimality result for the cross-validation selector among candidate estimators generated a new concept called “loss-based super learning,” which is a general system for fitting an infinite-dimensional parameter of the probability distribution of the data that allows one to map a very large library of candidate estimators into a new improved estimator. It made it clear that, given some global bounds on the semiparametric statistical model, humans should not choose the estimation procedure for fitting the probability distribution of the data, or a relevant part thereof, but an a priori defined estimator (i.e., the super learner) should fully utilize the data to make sound informed choices based on cross-validation. That is, the theory of super learning allows us to build machines that remove human intervention as much as possible.

Even though the theory teaches us that the super learner of the probability distribution does make the optimal bias–variance tradeoff with respect to the prob-

ability distribution as a whole (i.e., with respect to the dissimilarity between the super learner and the truth, as implied by the loss function), it is too biased for low-dimensional target features of the probability distribution, such as an effect of a variable/treatment/exposure on an outcome. The super learner is instructed to do well estimating the probability distribution, but the super learner was not told that it was going to be used to evaluate a one-dimensional feature of the probability distribution such as an effect of a treatment. As a consequence, the substitution estimator of a target parameter obtained by plugging in the super learner into the target parameter mapping is too biased.

By definition of an efficient estimator, it was clear that the efficient influence curve needed to play a role for these substitution estimators to become less biased and thereby asymptotically linear and efficient estimators of the target parameter. But how? The current literature on efficient estimation had used the efficient influence curve as an estimating function (van der Laan and Robins 2003), and one either completely solved the corresponding estimating equation or one used the first step of the Newton–Raphson method for solving the estimating equation (e.g., Bickel et al. 1997) in case one already had a root- $n$ -consistent initial estimator available. A new way of utilizing the efficient influence curve within the framework of loss-based learning needed to be determined.

The super learner had to be modified so that its excess bias was removed. The idea of the two-stage targeted maximum likelihood estimator was born: (1) use, for example, the super learner as the initial estimator, (2) propose a clever parametric statistical working model through the super learner, providing a family of candidate fluctuations of the super learner and treating the super learner as fixed offset, (3) choose the fluctuation that maximizes the likelihood (or whatever loss function was used for the super learner), and (4) iterate so that the resulting modified super learner solves the efficient influence curve estimating equation. This resulted in the original TMLE paper with Daniel B. Rubin (van der Laan and Rubin 2006), which provides a general recipe for defining a TMLE for any given data structure, semiparametric statistical model for the probability distribution, and target parameter mapping, and thereby served as the basis of this book.

TMLEs can also be represented as loss-based learning. Here, the loss function is defined as a targeted version of the loss function used by the initial estimator, where the nuisance parameter of this targeted loss function plays the role of the unknown fluctuation parameters in the TMLE steps. TMLEs are a special case of loss-based learning.

TMLEs solved the above mentioned remaining issues that the estimating equation methodology suffered from: a TMLE does not require that the efficient influence curve be an estimating function, a TMLE solves the efficient influence curve estimating equation but is not defined by it (just like a maximum likelihood estimator solves a score equation but is uniquely defined as a maximum of the log-likelihood), a TMLE is a substitution estimator and thus respects the global constraints of the statistical model, a TMLE naturally integrates loss-based super learning (i.e., generalized machine learning based on cross-validation) and can utilize the same loss function to select among different TMLEs indexed by different nuisance parameter

estimators that are needed to carry out the targeting update step. That is, even the choice of nuisance parameter estimator can now be tailored toward the target parameter of interest (van der Laan and Gruber 2010). Finally, under conditions that allow efficient estimation of the target parameter, a TMLE is an asymptotically efficient substitution estimator.  $\square$

**Sherri:** My methodological contributions have largely focused on adapting TMLE for case-control studies. Additionally, I've spent significant time with Mark formulating a general framework for teaching TMLE, with comprehensive notation and language, in a way that is accessible for researchers and students in fields such as epidemiology.

I received my B.S. in statistics in 2005 with the goal of going to graduate school for a career in medical research. Thus, I thought this meant I would be an “applied” statistician using existing tools. Then I took one of Mark’s upper division courses during the first year of my Ph.D. program at UC Berkeley. Even though I didn’t immediately understand all of the technical aspects of what he was teaching, the concepts made complete sense. I contacted him and projects took off immediately.

My point in this addendum to Mark’s journey is that you need not be a fully trained theoretical statistician to start understanding and using these methods. The work is driven by real-world problems, and thus is immediately applicable in practice. It is *theoretical* because new methods needed to be developed based on efficiency theory, but it is also very *applied*. You see this in the many examples that permeate this text. We don’t present anything that isn’t based on a real data set that we’ve encountered. In short, this book is not meant to sit on a shelf.  $\square$

**The book:** The book itself also went through a journey of its own. We started seriously writing for the book in January 2010 and for many months went back and forth debating the level we were trying to target. Should we generate a textbook that was more like an epidemiology text and would be broadly accessible to a greater number of applied readers with less formal statistical training? Should we develop a purely theoretical text that would mostly be of interest to a certain subset of statisticians? Ultimately, we struck a level that is somewhere in between these two extremes. Since there is no other book on targeted learning, we could not escape the inclusion of statistical formalism. However, we also did not want to lose all accessibility for nontheoreticians.

This led to a book that begins with six chapters that should be generally readable by most applied researchers familiar with basic statistical concepts and traditional data analysis. That is not to say many topics won’t be new and challenging, but these chapters are peppered with intuition and explanations to help readers along. The book progresses to more challenging topics and data structures, and follows a recognizable pattern via a road map for targeted learning and the general description of each targeted estimator. Thus, applied readers less interested in *why* it works and more interested in implementation can tease out those parts. Yet, mathematicians

and theoretical statisticians will not get bored, as extensive rigor is included in many chapters, as well as a detailed appendix containing proofs and derivations.

Lastly, this book is unique in that it also contains wonderful contributions from multiple invited authors, yet it is not a traditional edited text. As the authors of *Targeted Learning*, we have spent significant time crafting and reworking each of the contributed chapters to have consistent style, content, format, and notation as well as a familiar road map. This yields a truly cohesive book that reads easily as one text. □

## Intended Readership

We imagine a vast number of readers will be graduate students and researchers in statistics, biostatistics, and mathematics. This book was also written with epidemiologists, medical doctors, social scientists, and other applied researchers in mind. The first six chapters of the book, which comprise Part I, are a complete introduction to super learning and TMLE, including related concepts necessary to understand and apply these methods. Part I is designed to be accessible on many levels, and chapters that deal with more advanced statistical concepts feature guides that direct the reader to key information if they'd rather skip certain details. Additionally, these chapters could easily be used for a one-semester introductory course. The remaining chapters can be digested in any order that is useful to the reader, although we attempted to order them according to ease and subject matter. Parts II–IX handle more complex data structures and topics, but applied researchers will immediately recognize these data problems from their own research (e.g., continuous outcomes, case-control studies, time-dependent covariates, HIV data structures).

## Outline

**Introduction.** The book begins with an introduction written by Richard J.C.M. Starman titled “Models, Inference, and Truth: Probabilistic Reasoning in the Information Era.” This introduction puts the present state of affairs in statistical data analysis in a historical and philosophical perspective for the purpose of clarifying, understanding, and accounting for the current situation and to underline the relevance of topics addressed by TMLE for both the philosophy of statistics and the epistemology/philosophy of science. It identifies three major developments in the history of ideas that provide a context for the emergence of the probabilistic revolution and it discusses some important immanent developments in the history of statistics that have led to the current situation or at least may help to understand it.



## Part I – Targeted Learning: The Basics

The chapters in Part I of the book can stand alone as material for a complete introductory course on super learning and TMLE in realistic nonparametric and semi-parametric models. They cover essential information crucial to understanding this methodology, encapsulated in the convenient road map for targeted learning. We present in detail the TMLE of an additive causal effect of treatment on a binary or continuous outcome based on observing  $n$  independent and identically distributed random variables defined by the following type of experiment: randomly sample a subject from a population, measure baseline covariates, subsequently assign a treatment, and finally measure an outcome of interest. This TMLE is demonstrated in the estimation of the effect of vigorous exercise on survival in an elderly cohort.

**Chapter 1.** This chapter introduces the open problem of targeted statistical learning. We discuss, in general terms, the traditional approach to effect estimation as well as the concepts of data, data-generating distribution, model, and the target parameter of the data-generating distribution. We also motivate the need for estimators that are targeted and present the road map for targeted learning that will be explained in depth in Chaps. 2–5.

**Chapter 2.** In this chapter, readers will learn about structural causal models (SCMs), causal graphs, causal assumptions, counterfactuals, identifiability of the target parameter, and interpretations of the target parameter (i.e., causal or purely statistical). This material is essential background before moving on to the estimation steps in the road map for targeted learning. The chapter is based on the methods pioneered by Judea Pearl and are given thorough treatment in the recently published second edition of *Causality* (Pearl 2009).

**Chapter 3.** The first step in the TMLE is an initial estimate of the data-generating distribution  $P_0$ , or the relevant part  $Q_0$  of  $P_0$  that is needed to evaluate the target parameter. Estimation of  $Q_0$  incorporating the flexible ensemble learner super learner is presented in this chapter. Cross-validation is an essential component of super learning and is also presented. Simulation studies and multiple data analysis examples illustrate the advantages of super learning.

**Chapters 4 and 5.** In these two chapters, the TMLE methodology is presented in detail, including a conceptual overview, implementation, and theory. TMLE is a two-step procedure where one first obtains an estimate of the relevant portion  $Q_0$  of  $P_0$ . The second stage updates this initial fit in a step targeted toward making an optimal bias–variance tradeoff for the parameter of interest (i.e., target parameter), instead of the overall density  $P_0$ . It does this by proposing a parametric submodel through the initial fit of  $Q_0$ , and estimating the unknown parameter of this submodel that represents the amount of fluctuation of the initial fit. The submodel typically depends on a fit of a nuisance parameter such as a treatment or censoring mechanism. Finally, one evaluates the target parameter of this TMLE fit of  $Q_0$ , which is called the TMLE of the target parameter. The TMLE of the target parameter is double robust and can incorporate data-adaptive likelihood-based estimation procedures to estimate  $Q_0$  and the nuisance parameter. Inference (i.e., confidence intervals) and interpretation are also explained, concluding the road map for targeted learning.

**Chapter 6.** The many attractive properties of TMLE include the fact that it produces well-defined, loss-based, consistent, efficient substitution estimators of the target parameter. These topics are explained in depth, and the TMLE is compared to other estimators of a target parameter of the data-generating distribution, with respect to these properties.

## Part II – Additional Core Topics

Part II delves deeper into some core topics: the choice of submodel and loss function that defines the TMLE, causal parameters defined by marginal structural working models, and an in-depth coverage of methods dealing with violations of the positivity assumption. It focuses on experiments involving the measurement of baseline covariates, a treatment, possibly an intermediate random variable, and a final outcome.

**Chapter 7.** The TMLE of a parameter of a data-generating distribution, known to be an element of a semiparametric statistical model, involves constructing a parametric statistical working model through an initial density estimator with parameter  $\epsilon$  representing an amount of fluctuation of the initial density estimator, where the score of this fluctuation model at  $\epsilon = 0$  equals or spans the efficient influence curve/canonical gradient. The latter constraint can be satisfied by many parametric fluctuation models, since it represents only a local constraint of its behavior at zero fluctuation. However, it is very important that the fluctuations stay within the semiparametric statistical model for the observed data distribution, even if the parameter can be defined on fluctuations that fall outside the assumed observed data model. In particular, in the context of sparse data, a violation of this property can heavily affect the performance of the estimator. We demonstrate this in the context of estimation of a causal effect of a binary treatment on a continuous outcome that is bounded. It results in a TMLE that inherently respects known bounds and, consequently, is more robust in sparse data situations than the TMLE using a naive fluctuation model. The TMLE is based on a quasi-log-likelihood loss function and a logistic regression fluctuation model.

**Chapter 8.** In this chapter we consider estimation of a direct effect of treatment on an outcome in the presence of an intermediate variable. The causal model, the direct effect, the estimand defined by the identifiability result for the direct effect, and the TMLE of the target parameter are presented. As an illustration we estimate the direct effect of gender on salary in a gender-inequality study. It is shown that the same TMLE can be used to estimate the estimand defined by the identifiability result for the causal effect of a treatment on an outcome among the treated within an appropriate (different) causal model.

**Chapter 9.** One is often interested in assessing how the effect of a treatment is modified by some baseline covariates. For this purpose, we present marginal structural models that model the causal effect of treatment as a function of such effect modifiers. The TMLE of the unknown coefficients in the marginal structural model is presented. The marginal structural models are used as working models to define

the desired effect modification parameters, so that they do not make unrealistic assumptions in the causal model and thereby on the data-generating distribution. As an example, we assess the effect of missing doses on virologic failure as a function of the number of months of past viral suppression in an HIV cohort.

**Chapter 10.** The estimand that is defined by the identifiability result for the causal quantity of interest defines the target parameter of the data-generating distribution. The definition of the estimand itself often requires a particular support condition, which is called the positivity assumption. For example, the estimand that defines the additive causal effect of treatment on an outcome is only defined if for each value of the covariates (representing the confounders) there is a positive probability on both treatment and control. This chapter provides an in-depth discussion of the positivity assumption, and the detrimental effect of the practical or theoretical violation of this assumption on the statistical inference, due to the sparse-data bias induced by this violation. In addition, this chapter presents a parametric bootstrap-based diagnostic tool that allows one to diagnose this sparse-data bias. Its performance is demonstrated on simulated data sets and in assessing the effect of a mutation in the HIV virus on drug resistance in an HIV data application. Finally, the chapter presents common approaches to dealing with positivity violations and concludes with the presentation of a systematic general approach.

### **Part III – TMLE and Parametric Regression in Randomized Controlled Trials**

Part III still considers an experiment that generates baseline covariates, treatment, and a final outcome, as highlighted in Parts I and II, but it delves deeper into the special case where treatment is randomized. In this case, the TMLE is always consistent and asymptotically linear, thereby allowing the robust utilization of covariate information. We demonstrate that a TMLE that uses as initial estimator a maximum likelihood estimator according to a parametric regression model does not update the initial estimator, proving a remarkable robustness property of maximum likelihood estimation in randomized controlled trials (RCTs). In addition, we show how the fit of the parametric regression model (i.e., the initial estimator in the TMLE) can be optimized with respect to the asymptotic variance of the resulting TMLE, thereby guaranteeing improvement over existing practice.

**Chapter 11.** The TMLE of a causal effect of treatment on a continuous or binary outcome in an RCT is presented. It is shown that the TMLE can be based on a maximum likelihood estimator according to a generalized linear working model, where the maximum likelihood estimation fit is inputted in the target parameter mapping defined by the so-called g-formula for the desired causal effect.

**Chapter 12.** As in Chap. 11, the TMLE in this chapter is based on a parametric regression model, but the coefficients of the initial estimator in the TMLE are fitted so that the resulting TMLE has minimal asymptotic variance. This results in a TMLE that is guaranteed to outperform current practice (i.e., unadjusted estimator), even if the parametric model is heavily misspecified. Other estimators presented in

the literature are also discussed, and a simulation study is used to evaluate the small sample performance of these estimators.

## Part IV – Case-Control Studies

The data-generating experiment now involves an additional complexity called biased sampling. That is, one assumes the underlying experiment that randomly samples a unit from a target population, measures baseline characteristics, assigns a treatment/exposure, and measures a final binary outcome, but one samples from the conditional probability distribution, given the value of the binary outcome. One still wishes to assess the causal effect of treatment on the binary outcome for the target population. The TMLE of a causal effect of treatment on the binary outcome based on such case-control studies is presented. Matched case-control studies are considered as well. It is also shown how to apply super learning to risk prediction in a nested case-control study.

**Chapter 13.** Case-control study designs are frequently used in public health and medical research to assess potential risk factors for disease. These study designs are particularly attractive to investigators researching rare diseases, as they are able to sample known cases of disease, vs. following a large number of subjects and waiting for disease onset in a relatively small number of individuals. Our proposed case-control-weighted TMLE for case-control studies relies on knowledge of the true prevalence probability, or a reasonable estimate of this probability, to eliminate the bias of the case-control sampling design. We use the prevalence probability in case-control weights, and our case-control weighting scheme successfully maps the TMLE for a random sample into a method for case-control sampling.

**Chapter 14.** Individually matched case-control study designs are commonly implemented in the field of public health. While matching is intended to eliminate confounding, the main *potential* benefit of matching in case-control studies is a gain in efficiency. This chapter investigates the use of the case-control-weighted TMLE to estimate causal effects in matched case-control study designs. We compare the case-control-weighted TMLE in matched and unmatched designs in an effort to determine which design yields the most information about the causal effect. In many practical situations where a causal effect is the parameter of interest, researchers may be better served using an unmatched design.

**Chapter 15.** Using nested case-control data from a large Kaiser Permanente database, we generate a function for mortality risk prediction with super learning. The ensemble super learner for predicting death (risk score) outperformed all single algorithms in the collection of algorithms, although its performance was similar to several included algorithms. Super learner improved upon the worst algorithms by 17% with respect to estimated risk.

## Part V – RCTs with Survival Outcomes

In Part V we consider the following experiment: one randomly samples a unit from a target population, measures baseline characteristics, randomly assigns a treatment, and follows the subject to the minimum of dropout, the time to event of interest, and time to the end of study. The dropout time is allowed to be affected by the baseline covariates. We present the TMLE of the causal effect of treatment on survival, and we also consider effect modification by discrete baseline factors.

**Chapter 16.** In most RCTs, the primary outcome is a time-to-event outcome that may not be observed due to dropout or end of follow-up. The dropout or right censoring time may depend on the baseline characteristics of the study subject. The TMLE of a causal effect of treatment on the survival function of such a time-to-event outcome requires estimation of the conditional failure time hazard as a function of time, treatment, and the baseline covariates. The super learner of this hazard function is presented and is demonstrated with a lung cancer RCT.

**Chapter 17.** The TMLE of a causal effect of treatment on a survival function in an RCT is presented. This requires an update of the initial estimator of the conditional hazard function (e.g., super learner), where the update relies on an estimator of the right censoring mechanism and the treatment assignment mechanism (where the latter is known in an RCT). The statistical properties of the TMLE are discussed showing that it provides a superior alternative to current practice in terms of unadjusted Cox proportional hazards estimators or multiple imputation (maximum likelihood estimation)-based estimators.

**Chapter 18.** It is often of interest to assess if the causal effect of treatment on survival is modified by some baseline factors. In this chapter, we define the appropriate causal model and the target parameters that quantify effect modification by a discrete baseline factor. We present the TMLE of these effect modification parameters. The TMLE is demonstrated on an HIV clinical trial to assess effect modification by gender and by baseline CD4 in an HIV study. The results are contrasted with current practice, demonstrating the great utility of targeted learning.

## Part VI – C-TMLE

Collaborative TMLE (C-TMLE) provides a further advance within the framework of TMLE by tailoring the fit of the nuisance parameter required in the TMLE-step for the purpose of the resulting TMLE of the target parameter. That is, the C-TMLE introduces another level of targeting beyond a regular TMLE. This part demonstrates the C-TMLE for the causal effect of treatment on an outcome, including time-to-event outcomes that are subject to right censoring. Simulation studies as well as data analyses are provided to demonstrate the practical utility of C-TMLE.

**Chapter 19.** The C-TMLE of the additive causal effect of treatment on an outcome is presented, allowing an a priori-specified algorithm to decide what covariates to include in the treatment mechanism fit, where the decisions are based on a loss-function that measures the fit of the corresponding TMLE instead of the fit of the

treatment mechanism itself. The TMLE and C-TMLE are compared in simulation studies. The C-TMLE is also applied to assess the effect of all mutations in the HIV virus on drug-resistance, controlling for the history of the patient, dealing with the many strong correlations between mutations resulting in practical violations of the positivity assumption.

**Chapter 20.** The C-TMLE of the causal effect of treatment on a survival time that is subject to right censoring is developed. A simulation study is used to evaluate its practical performance in the context of different degrees of violation of the positivity assumption.

**Chapter 21.** This chapter uses simulation studies proposed in the literature to evaluate a variety of estimators for estimating the mean of an outcome under missingness, and the additive effect of treatment when treatment is affected (i.e., confounded) by baseline covariates. These simulations are tailored to result in serious practical violations of the positivity assumption, causing a lot of instability and challenges for double robust efficient estimators such as the TMLE. These simulations have been extensively debated in the literature. This chapter includes TMLE and C-TMLE in the debate. We contrast the C-TMLE to the TMLE and other estimators, showing that the C-TMLE is able to deal with sparsity (i.e., violations of positivity) in a sensible and robust way, while still preserving the optimal asymptotic properties of TMLE.

## Part VII – Genomics

In Part VII we consider the experiment in which one randomly samples a unit from a target population, one measures a whole genomic profile on the unit, beyond other baseline characteristics, one possibly measures a treatment, and one measures a final outcome. In such studies one is often interested in assessing the effect of each genomic variable on the outcome or on the effect of the treatment. TMLE targets the effect of each genomic variable separately, contrary to current practice in variable importance analysis. These genomic variables are often continuous, so that one needs to define an effect of a continuous marker on the outcome of interest. For that purpose we employ semiparametric regression models. The TMLE of the effect measures defined by these semiparametric regression models are presented, and demonstrated in genomic data analyses.

**Chapter 22.** The TMLE for assessing the effect of biomarkers is presented and compared with other methods for variable importance analysis, such as random forest, in a comprehensive simulation study, and a breast cancer gene expression study.

**Chapter 23.** We present the TMLE and C-TMLE for assessing the effect of a marker on a quantitative trait, across a very large number of markers along the whole genome. Simulations and genomic data analyses are used to demonstrate the TMLE and C-TMLE.

## Part VIII – Longitudinal Data Structures

In Part VI, we consider experiments that generate the full complexity of current day longitudinal data structures: one randomly samples a unit from a target population, measures baseline characteristics, and at regular or irregular monitoring times collects measurements on time-dependent treatments or exposures, time-dependent covariates, and intermediate outcomes, until the minimum of right-censoring or time to the event of interest. Observing such longitudinal data structures on a unit allows the identification of causal effects of multiple time point treatment regimens as well as individualized treatment rules. In this part, we demonstrate the roadmap for addressing the scientific questions of interest and the corresponding TMLE for three such longitudinal case studies. Technically-inclined readers may first wish to read the longitudinal sections of Appendix A before digesting these chapters.

**Chapter 24.** A longitudinal HIV cohort is presented and three scientific questions of interest are formulated. The road map is applied. It starts out with the definition of the causal model, the definition of the target causal parameters that represent the answers to the scientific questions, and the identifiability result resulting in the estimand of interest. The statistical model and the estimand/target parameter of the data-generating distribution define the estimation problem. Different methods for estimation are reviewed and presented: maximum likelihood estimation, inverse probability of censoring weighted estimation (IPCW), targeted maximum likelihood estimation, and inefficient practically appealing TMLEs referred to as IPCW reduced-data TMLEs.

**Chapter 25.** A longitudinal study is presented which involves the follow up of women going through an in vitro fertilization (IVF) program. One is interested in assessing the probability of success of a complete IVF program. The road map is applied as in all chapters. The TMLE of the probability of success of a complete IVF program is developed, and applied to the study. Simulations are also presented.

**Chapter 26.** In this chapter, targeted maximum likelihood learning is illustrated with a data analysis from a longitudinal observational study to investigate the question of “when to start” antiretroviral therapy to reduce the incidence of AIDS defining cancer in a population of HIV infected patients. Two treatment rules are considered: (1) start when CD4 count drops below 350, and (2) start when CD4 count drops below 200. The TMLE of the corresponding causal contrast is developed and applied to the database maintained by Kaiser Permanente.

## Part IX – Advanced Topics

We deal with the following explicit questions. Is the utilization of machine learning in the TMLE a concern for establishing asymptotic normality? Can we develop a TMLE for group sequential adaptive designs in which the treatment assignment probabilities are set in response to the data collected in previously observed groups? What are the asymptotics of this TMLE for such a complex experiment in which all subjects are correlated due to treatment assignment being a function of the outcomes

of previously observed subjects? Does sequential testing still apply? Since Bayesian learning is nontargeted and suffers from the same drawbacks as maximum likelihood based estimation, can we employ the principles of TMLE to construct a targeted Bayesian learning method?

**Chapter 27.** The cross-validated TMLE (CV-TMLE) is presented where asymptotic linearity and efficiency can be established under minimal conditions. A formal theorem is presented for the CV-TMLE of the additive causal effect, demonstrating that it is able to fully utilize all the machine learning power in the world while still allowing, and, in fact, enhancing, valid statistical inference.

**Chapter 28.** It is shown that the TMLE procedure naturally lends itself to targeted Bayesian learning in which a prior probability distribution on the target parameter of interest is mapped into a posterior distribution of the target parameter of interest. The frequentist properties of the mean and spread of the posterior distribution are established showing that the proposed procedure is completely valid: the mean of the posterior distribution is a double robust efficient estimator of the target parameter, and the posterior distribution yields valid credible intervals.

**Chapter 29.** We consider targeted group sequential adaptive designs that adapt the randomization probabilities in response to all the data collected in previous stages. The TMLE of the desired causal effect of the treatment is developed and presented. Asymptotics of the TMLE are based on martingale central limit theorems. It is shown that sequential testing can still be naturally embedded in such adaptive group sequential designs.

## Part X – Appendices

Part X consists of two appendices providing important supplementary material in support of the central text. The core of the first appendix is a theoretical guide covering essential topics, derivations, and proofs. This is followed by a brief introduction to R code for super learning and TMLE. Additional R code is available on the book's website: [www.targetedlearningbook.com](http://www.targetedlearningbook.com).

**Appendix A.** This appendix provides a succinct but comprehensive review of the empirical process, asymptotic linearity, influence curves, and efficiency theory. This theory establishes the theoretical underpinnings of TMLE, C-TMLE, and CV-TMLE. In addition, Appendix A provides a generic approach that allows one to compute a TMLE on a new estimation problem in terms of the definition of the data structure, data-generating distribution, the statistical model, and the target parameter mapping that maps a probability distribution in its target parameter value. The TMLE for general longitudinal data structures is presented. A variety of examples are used to demonstrate the power of this generic machinery for computing a TMLE. Appendix A can be used to teach an advanced class about the theory of estimation and, in particular, of TMLE.

**Appendix B.** This brief appendix provides R code and links to R code for each of the implementations of the TMLE as presented in this book.



## Acknowledgements

This book reflects the greater effort of a large team of people, without whom it would not have been completed. We thank the following students for their work on these methods: Laura Balzer, Katie Benton, Jordan Brooks, Paul Chaffee, Iván Díaz Muñoz, Susan Gruber, Sam Lendle, Kristin E. Porter, Stephanie Sapp, Boriska Toth, Catherine Tuglus, and Wenjing Zheng. We also thank our colleagues, many of whom contributed a chapter in this book, for their contributions to and discussions on these methods, as well as their vital collaborative efforts: Ben Arnold (UC Berkeley), David R. Bangsberg (Harvard), Oliver Bembom (AdBrite), Marco Carone (UC Berkeley), Antoine Chambaz (Université Paris Descartes), Jack Colford (UC Berkeley), Steven G. Deeks (UCSF), Victor De Gruttola (Harvard), W. Jeffrey Fessel (Kaiser Permanente), Bruce Fireman (Kaiser Permanente), Alan E. Hubbard (UC Berkeley), Nicholas P. Jewell (UC Berkeley), Kelly L. Moore (Jewish Home), Torsten B. Neilands (UCSF), Romain S. Neugebauer (Kaiser Permanente), Michelle Odden (UCSF), Judea Pearl (UCLA), Maya L. Petersen (UC Berkeley), Eric C. Polley (NCI), Art Reingold (UC Berkeley), Elise D. Riley (UCSF), Michael Rosenblum (Johns Hopkins), Daniel B. Rubin (FDA), Robert W. Shafer (Stanford), Jasjeet S. Sekhon (UC Berkeley), Michael J. Silverberg (Kaiser Permanente), Greg Soon (FDA), Richard J.C.M. Starmans (Universiteit Utrecht), Ori M. Stitelman (UC Berkeley), Ira B. Tager (UC Berkeley), Thamban Valappil (FDA), Hui Wang (Stanford), Yue Wang (Novartis), and C. William Wester (Vanderbilt).

Much of the content presented in this book has been taught by Mark in multiple upper-division biostatistics courses at UC Berkeley (e.g., PH240B, PH243D, PH243A, and PHC246A). The authors would like to thank the students in these courses, as their questions guided the format of this text. We thank Maya L. Petersen and Ira B. Tager for many insightful discussions on content and organization for Part I of the book. We also thank our thorough and patient copyeditor Glenn Corey, who took our manuscript to the next level. Computing is a central component of the work we do, and thus we thank our systems administrator Burke J. Bundy.

Berkeley, January 2011

*Mark J. van der Laan  
Sherri Rose*

Targeted Learning

Causal Inference for Observational and Experimental  
Data

van der Laan, M.J.; Rose, S.

2011, LXXII, 628 p., Hardcover

ISBN: 978-1-4419-9781-4