*University of California, Berkeley*

U.C. Berkeley Division of Biostatistics Working Paper Series

*Year* 2012 *Paper* 297

# Targeted Learning for Causality and Statistical Analysis in Medical Research

Sherri Rose* Richard J.C.M. Starmans†

Mark J. van der Laan‡

*Johns Hopkins Bloomberg School of Public Health, sherrirosephd@gmail.com

†Universiteit Utrecht, starmans@cs.uu.nl

‡School of Public Health, Division of Biostatistics, University of California, Berkeley, laan@berkeley.edu



http://biostats.bepress.com/ucbbiostat/paper297

# Targeted Learning for Causality and Statistical Analysis in Medical Research

Sherri Rose, Richard J.C.M. Starmans, and Mark J. van der Laan

**Abstract**

The authors present the use of targeted learning methods for medical research, prepared as a chapter for the upcoming book "Statistics: Discovering Your Future Power." The targeted learning framework involves the explicit specification of the data, model, and parameter. The estimators are double robust and efficient, and can incorporate machine learning procedures such as the super learner.

# Targeted Learning for Causality and Statistical Analysis in Medical Research

a chapter for the textbook *Statistics: Discovering Your Future Power*

Sherri Rose[1], Richard J.C.M. Starmans[2], and Mark J. van der Laan[3]

[1] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
[2] Department of Information and Computing Sciences, Universiteit Utrecht
[3] Division of Biostatistics, University of California, Berkeley

**Abstract**

The authors present the use of targeted learning methods for medical research, prepared as a chapter for the upcoming book *Statistics: Discovering Your Future Power*. The targeted learning framework involves the explicit specification of the data, model, and parameter. The estimators are double robust and efficient, and can incorporate machine learning procedures such as the super learner.

## What is Targeted Learning?

When we ask scientific questions, we frequently collect data in an attempt to answer these questions. In the many areas of medical research, we are often interested in so-called causal effects. That is to say, we prefer not to merely conclude that there is an association or correlation between two variables. Instead, we want to know that variable $X$ causes variable $Y$.

Let's introduce a tangible example. Suppose we are studying the causal effect of treatment noncompliance on intestinal cancer among patients with celiac disease in the United States. Celiac disease is an autoimmune disorder characterized by the body's inability to digest the protein gluten, which is found in certain grains (wheat, barley, and rye). Gluten causes significant damage to the small intestines in people with celiac disease, and can cause intestinal cancer over time. The only available treatment for celiac disease is a completely gluten-free diet. Noncompliance on the diet is nontrivial, with an estimated 30% of those diagnosed with the disease purposely noncompliant. For simplicity, let's define the treatment as binary and either compliant (always follows the diet) or not compliant (sometimes or never follows the diet).

We've identified our population of interest, persons with celiac disease in the United States, and now we must decide how to design our study to answer the question of interest. It's not feasible for us to collect information about each person in the United States who has celiac disease. Thus, we must take a sample from our population of interest in order to make inferences about the population. Experimental studies in human populations regularly occur under the umbrella of randomized controlled trials (RCTs). Here, we assign a treatment or exposure in the experiment based on randomization.

In a randomized study of diet noncompliance in celiac disease on intestinal cancer where each subject was assigned to treatment based on tossing a fair coin, the differences between the compliant and noncompliant groups would be due only to the treatment since other factors are balanced,

up to random error. This would be a causal effect. However, in practice, randomization does not occur perfectly, and certain covariates may be predictive of the outcome. It is also important to note that treatment or exposure assignment can have varying degrees of complexity, and occur over time or at one intervention.

Just as in a hypothetical RCT studying smoking exposure, it would not be ethical to require celiac disease patients to ingest significant amounts of gluten for any lengthy period of time. The outcome (intestinal cancer) also takes many years to develop and an RCT would be cost prohibitive. Thus, many studies are "observational," where the treatment or exposure is not randomized, and we observe and record the subjects' behavior without assignment. Since our observational study of celiac disease does not involve randomization, differences between groups will be biased due to various issues such as confounding, informative missingness, and censoring.

Traditional analysis methods for observational studies involve assuming restrictive parametric statistical models and using maximum likelihood estimation. There are many problems with this nontargeted approach. A key issue is that we are estimating many coefficients (parameters) in a parametric statistical model, and we use a criterion that is focused on the fit of the entire probability distribution. However, we are truly interested in the causal effect of treatment noncompliance on intestinal cancer, not multiple coefficients in a statistical model. This is then clearly not ideal as we have spread the error across the whole distribution. If the statistical model is correct, the effect of the treatment is a conditional effect, and not necessarily the parameter we are interest in estimating. Additionally, when the statistical model is misspecified, the parameters in the statistical model are not interpretable. In most medical applications, we have many covariates and this bias can be substantial. With targeted learning, we isolate the parameter of interest directly in less restrictive and more realistic models. Our estimator has important statistical properties that improve on standard methods, and also allows the incorporation of aggressive and flexible machine learning approaches.

# Defining the Research Question

The first step in targeted learning is accurately defining the question of interest. This includes a clear description of the data, statistical model, and parameter. Our study is an experiment where we draw a random variable (an individual subject) from our celiac population $n$ times. The data we observe are realizations of these $n$ random variables, and the random variables have an underlying probability distribution. Now remember: a statistical model need not be a parametric statistical model! A statistical model in general represents the set of possible probability distributions of the data. Our statistical model should represent our knowledge about the data. In our celiac disease study, we will assume a nonparametric statistical model. We are assuming that we know the data are comprised of observations on $n$ independent and identically distributed random variables, which is a real assumption, but we make no other assumptions. A parametric statistical model would assume that the probability distribution underlying the data is known (up to a certain number of parameters). Our statistical model makes no such assumption, as, in practice, it is widely known that nonsaturated parametric statistical models are wrong.

Now, we've made only those assumptions in our nonparametric model that are supported by the data. But there is nothing about this statistical model that allows us to interpret our parame-

ter as causal...yet. We can make additional causal assumptions, and these assumptions combined with our statistical model are referred to simply as the model for the observed data. These causal assumptions allow us set up a system where we can intervene and "set" an individual to be compliant or not compliant on their gluten-free diet. It is this system, that we will define more formally shortly, that will enable us to define the causal effect of diet compliance on intestinal cancer among celiac patients in the United States.

We've just described in words what we are interested in: we want to estimate this causal effect. We now need to establish its identifiability. If the assumptions we have made based on background knowledge are enough to express the causal parameter as a parameter of the probability distribution of the observed data, and thereby prove its identifiability, we need not make additional causal assumptions. However, if this is not the case, we need additional assumptions for establishing identifiability of the desired causal effect. The identifiability result will allow us to define an estimand (i.e., statistical target parameter) as a function of the data generating distribution, which can be interpreted purely as a statistical target parameter, and, under these additional assumptions, it can also be interpreted as a causal effect. Either way, the definition of the statistical model and statistical target parameter defines our estimation problem, while the possible additional causal assumptions required to interpret the statistical target parameter as a causal effect only concern the interpretability of the statistical output.

### Data and Statistical Model

Let us formalize these concepts with some notation. The data consists of $n$ i.i.d. copies of random variable $O \sim P_0$, where $P_0$ is the true underlying probability distribution for $O$. In this chapter we will explore a simple case, where $O$ is defined as: $O = (W,A,Y) \sim P_0$. This data structure has baseline covariates $W$, exposure variable (gluten-free diet compliance) $A$, and outcome (intestinal cancer within 10 years of baseline) $Y$. The variable $W$ is a vector that includes age, sex, concomitant health risks, and other important potential confounding variables. This simplified example ignores issues of censoring, missingness, and that our exposure of interest might be time-dependent depending on our scientific question. The statistical model $\mathscr{M}$ is the set of possible probability distributions for $P_0$, is nonparametric, and incorporates our subject matter knowledge about $P_0$.

### Model with Causal Assumptions

The model includes additional causal assumptions, which allows us to interpret the parameter of interest as a causal effect. The causal assumptions are nontestable assumptions, as we cannot use the data to verify their accuracy. We will assume a so-called structural causal model (SCM), comprised of endogenous variables $X = (X_j : j)$ and exogenous variables $U = (U_{X_j} : j)$. The SCM describes that each $X_j$ is a deterministic function of other endogenous variables and an exogenous error $U_j$. The errors $U$ are never observed. For each $X_j$ we characterize its parents from among $X$ with $Pa(X_j)$. For example, in our celiac study, $X = (W,A,Y)$, and $Pa(A) = W$. We know this due to the time ordering of the variables. Thus we can now write:

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

and the functional form of $f_{X_j}$ is often unspecified. An SCM can be fully parametric, but it is not recommended as our background knowledge does not support the assumptions involved. The joint distribution of $U$ and the collection of functions $f_{X_j}$ indexed by the complete set of all endogenous variables $f = (f_{X_j} : j)$ described the data-generating distribution of $(U,X)$.

There is a corresponding SCM for $O$, and this involves defining the relationship between $O$ and $(U,X)$ so that the SCM for the full data suggests a parameterization of the probability distribution of $O$ in terms of the distribution $P_U$ of $U$ and $f$. The SCM for the observed data suggests a statistical model for the probability distribution of $O$.

To summarize, the causal assumptions we make with our SCM for the full data are that (1) each endogenous variable depends on other endogenous variables only through its parents and (2) our exogenous variables have a certain joint distribution $P_U$. The SCM for the observed data additionally assumes that the probability distribution of $O$ is suggested by the probability distribution of $(U,X)$.

Let's demonstrate these concepts in our celiac disease study. We have the functions $f = (f_W, f_A, f_Y)$ and exogenous variables $U = (U_W, U_A, U_Y)$. Our structural equation models, based on subject matter knowledge, are given as:

$$\begin{aligned} W &= f_W(U_W), \\ A &= f_A(W, U_A), \\ Y &= f_Y(W, A, U_Y). \end{aligned} \tag{1}$$

We have not made any assumptions about the form of the functions $f$. We do now make the assumption that the observed data structure $O = (W,A,Y)$ is a realization of the endogenous variables $(W,A,Y)$ produced by this system, which defines the SCM for the observed data.

We can also present a causal graph based on our SCM, which is a common graphic to describe some of the causal assumptions made by our SCM. Figure 1 displays two possible causal graphs that could be drawn based on the SCM given in equation (1), they have different assumptions about the distribution of $P_U$. Often we know nothing about this distribution, but for the identifiability of our causal parameter, we may be required to make additional assumptions. For example, we may need to assume in our SCM that $U_A$ is independent of $U_Y$, given $W$ in order for our causal effect to be identifiable. This is equivalent to assuming there are no unmeasured confounders.

We can then write $Y_a = f_Y(W,a,U_Y)$, which is as a random variable corresponding to intervention $A = a$ and the marginal probability distribution of $Y_a$ is given by $P_{U,X}(Y_a = y) = P_{U,X}(f_Y(W,a,U_Y) = y)$. The causal effect the risk difference in our study of diet compliance on intestinal cancer in celiac patients can now be defined as a parameter of the distribution of $(U,X)$ given by

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

This causal quantity is the difference in marginal means of counterfactuals $Y_1$ and $Y_0$.

### Target Parameter

We've already described what we are interested in learning in our study of celiac disease subjects in the United States: we want to understand the causal effect of gluten-free diet noncompliance on intestinal cancer. This difference can be measured on an additive scale in the form of a risk

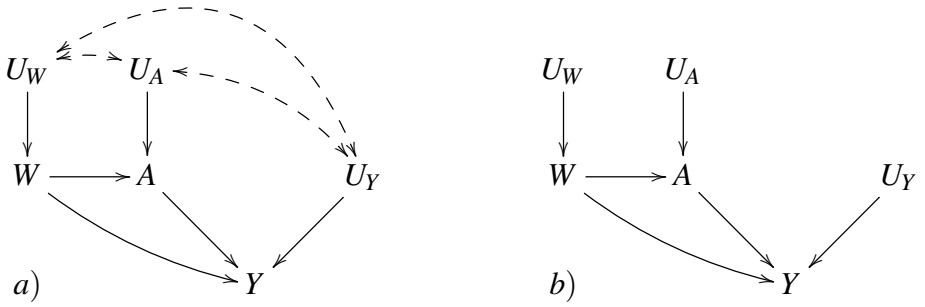


Figure 1: $a)$ A causal graph for (1) with no assumptions on the distribution of $P_U$. $b)$ A possible causal graph for (1) with all $U_j$ independent.

difference or on a multiplicative scale, such as a the familiar relative risk or odds ratio. There are also more complicated parameters that may be of interest depending on your research question. Perhaps you'd like to stratify by gender, for example. We will present the risk difference for our study.

Suppose now we've made enough assumptions in our model that our causal quantity is identifiable, therefore we can write the causal parameter $\Psi^F(P_{U,X})$ as a parameter $\Psi(P)$ of an observed data probability distribution $P = P_{P_{U,X}}$ of $O$ implied by a distribution $P_{U,X}$ of the underlying $U,X$. $\Psi$ maps the true data distribution $P_0$ into our feature of interest. That is, the parameter $\Psi(P_0)$ is a function of the unknown distribution $P_0$. Remember, we are not simply grabbing a coefficient from a parametric model, we need to think carefully about the feature of the probability distribution we are interested in understanding.

Under the randomization assumption $U_A \perp U_Y$, conditional on $W$, and the positivity assumption $0 < P_0(A = 1 \mid W) < 1$, the causal additive risk difference can be defined as the following function of $P_0$ of $O$:

$$\Psi(P_0) = E_{W,0}[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W)], \tag{2}$$

where $E_0(Y \mid A = a, W)$ is the conditional mean of $Y$ given $A = a$ and $W$. Recall that $A$, diet compliance, is binary. $\Psi(P_0)$ can also be written as:

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

where

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

We will use a targeted maximum likelihood estimator (TMLE) to produce an estimate of $\Psi(P_0)$. A confidence interval will also be generated using either influence curve based methods. Interpretation is the last step, and we have two options. One is as a purely statistical parameter of $P_0$, and one as a causal parameter under additional (causal) assumptions (i.e, randomization and positivity assumption).

# Super Machine Learning

The first step in estimating our parameter of interest, given in (2), using TMLE, is a starting estimate of the piece of $P_0$ necessary to evaluate our target parameter. We want to use an estimator that respects our nonparametric model, one that acknowledges that $P_0$ is known only as an element of our nonparametric model. Our parameter depends on $P_0$ through the conditional mean $\bar{Q}_0(A,W) = E_0(Y \mid A,W)$ and the marginal distribution $Q_{W,0}$ of $W$. Thus, we can also denote our parameter as $\Psi(Q_0)$, where $Q_0 = (\bar{Q}_0, Q_{W,0})$. We will estimate the expectation over our covariates $W$ with the empirical mean over $W_i$, $i = 1,\ldots,n$. Therefore, $\bar{Q}_0(A,W) = E_0(Y \mid A,W)$ is the object we still need a flexible nonparametric way to estimate. Our task is then clear, the first step of the TMLE for the risk difference is to obtain an estimate of the conditional mean function $\bar{Q}_0(A,W)$. Our "plug-in" TMLE is of the type

$$\psi_n = \Psi(Q_n) = \frac{1}{n}\sum_{i=1}^{n}\{\bar{Q}_n(1,W_i) - \bar{Q}_n(0,W_i)\}. \tag{3}$$

The estimate is produced by plugging $Q_n = (\bar{Q}_n, Q_{W,n})$ into the parameter mapping $\Psi$.

The main knowledge we have about $P_0$ is that the data are $n$ i.i.d. realizations $o_1,\ldots,o_n$ on $n$ i.i.d. copies $O_1,\ldots,O_n$ of $O \sim P_0$. Since our outcome (intestinal cancer) is binary, our we are interested in estimating $\bar{Q}_0(A,W) = P_0(Y = 1 \mid A,W)$ as the first step of our TMLE. We do not merely want to use a parametric regression to estimate $\bar{Q}_0(A,W)$ as it makes very restrictive assumptions we cannot support with our background knowledge. This would lead to bias, bias we cannot remove by increasing our sample size. We want to use very flexible methods to produced an automated algorithm to estimate $\bar{Q}_0(A,W)$. Machine learning, or data-adaptive, methods also aim to "smooth" our data, but they do so less rigidly than a parametric regression. But which algorithm do we use? The literature has demonstrated that some methods will work best in certain data, while others will perform better in the next data set. It's unpredictable. How do we handle the problem of many algorithms and varied performance?

### Loss Function

Before we can figure out what estimator is the best estimator of $\bar{Q}_0(A,W)$, we must have a metric to define what we mean by "best." A loss function is the way we do this. A loss function $L$ assigns a performance measure to a candidate function $\bar{Q}$ applied to an observation $O$: $L : (O,\bar{Q}) \to L(O,\bar{Q}) \in \mathbb{R}$. For our study, we can use the $L_2$ squared error loss function $L(O,\bar{Q}) = (Y - \bar{Q}(A,W))^2$. or the log-likelihood loss $L(O,\bar{Q}) = -\{Y \log \bar{Q}(A,W) + (1-Y)\log(1-\bar{Q}(A,W))\}$. Since we have a binary $Y$, the $L_2$ loss indeed targets the function $\bar{Q}_0(A,W) = P_0(Y = 1 \mid A,W)$.

Our parameter of interest $\bar{Q}_0(A,W) = E_0(Y \mid A,W)$ is therefore the minimizer of the expected squared error loss or expected log-likelihood loss: $\bar{Q}_0 = \arg\min_{\bar{Q}} E_0 L(O,\bar{Q})$. $E_0 L(O,\bar{Q})$ evaluates $\bar{Q}$, and it is minimized at the optimal $\bar{Q}_0$. The expected loss is also referred to as the risk. This is the metric we will use to select the optimal algorithm, we want the estimator of the function $\bar{Q}_0$ with the realized value that minimizes the risk. The next question is: how do we know which algorithm from our collection of algorithms has the best performance with respect to this dissimilarity measure?

**Super Learner**

We know there are many different algorithms available to produce an estimate of $\bar{Q}_0$. We want the best algorithm that minimizes our risk, but how to we use many algorithms honestly to generate the best algorithm? In summary, ensembling allows us to use many algorithms to generate an ideal prediction function that is a weighted average of all the algorithms. Our final choice of weighted average is based on our a priori-selected criterion and we will use cross-validation methods to avoid a biased final prediction function. We will describe the use of ensembling in the super learner.

First, let's input our fictional celiac data set and a collection of potential algorithms. This collection could include multiple different regressions, random forests, support vector machines, or other algorithms. Let's divide our fictional celiac data set into $V$ mutually exclusive and exhaustive sets of size $\sim n/V$. A common choice of $V$ is 10, and we will perform 10-fold cross-validation within our ensembling super learner algorithm. Within each fold, 1 block is designated the "validation" set and we fit each of the algorithms on the remaining data called the training set. For example, in Fold 1, Block 1 is the validation set and we fit each algorithm on Blocks 2 through 10. In Fold 2, Block 2 is the validation set and we fit each algorithm on Block 1 and Blocks 3 through 10. This is done for each of the 10 folds such that we have 10 fits for each algorithm, one for every fold.

The next step involves generating predicted probabilities of intestinal cancer for each algorithm using the validation sets, based on the corresponding training set fit. Since each subject appears in only one validation set, at the end of this step we have one predicted probability for each subject per algorithm. An estimated risk is then calculated for each algorithm using the predicted probabilities generated with the cross-validation. This leaves us with one estimated cross-validated risk for each algorithm. We could stop here, and simply choose the algorithm with the smallest risk, but we can actually improve on the algorithm (called the *discrete* super learner) that does this.

A super learner builds a library of algorithms that consists of all possible weighted averages of the algorithms we input at the start of our super learner. This is now an augmented library, an even larger collection of algorithms, and one of the weighted averages may perform better any of the single algorithms alone. We apply the same cross-validation selector we used in the discrete super learner to this augmented set of candidate algorithms to produce the super learner. While we are minimizing the cross-validated risk over this infinite set of weighted averages, this procedure is actually not computationally infeasible, and requires a comparatively trivial calculation of an optimal weight vector.

We left our algorithm where we had cross-validated risks for each algorithm in our original collection of algorithms. We now propose a family of weighted combinations of these algorithms, which is indexed by a weight vector $\alpha$. We limit the vector $\alpha$ to include only those vectors that sum to one and where each weight is $\geq 0$. The optimal weight vector is selected in a simple minimization problem: a regression of $Y$ on the predicted probabilities of intestinal cancer. The weight vector selected is the one that minimizes the estimated expected squared error loss function. The last step involves fitting each algorithm on the complete celiac data set and combining these fits with the optimal weight vector. This is the super learner predictor function is a new estimator which can be used to generate predicted values for $\bar{Q}_0(A,W)$. To calculate an estimated risk for the super learner, the entire super learner algorithm is cross-validated.

In realistic scenarios (where a correctly specified parametric regression is not among the candi-

date algorithms), as long as our loss function is bounded, the super learner will perform as well as the so-called oracle selector, for large samples. The oracle is the algorithm in the library that is the closest to $\bar{Q}_0$, and recall that with the super learner, our algorithm library is the family of weighted combinations of estimators. When the library does include a correctly specified parametric regression, the super learner approximates the truth as fast as the parametric regression, but it is more variable.

# Targeted Maximum Likelihood Estimator

We have an initial super learner estimate of the relevant part of the data-generating distribution obtained, thus we can present the remainder of the TMLE procedure. This second stage updates our initial fit in a step targeted towards making an optimal bias–variance tradeoff for the parameter of interest, instead of the overall probability distribution. This leads to a targeted estimator of $Q_0$, and thereby in the corresponding substitution estimator of $\Psi(Q_0)$.

TMLE has several key statistical properties. The first is double robustness. TMLE removes the asymptotic bias of the initial estimator if it uses a consistent estimator of the treatment mechanism $P_0(A \mid W)$. Even when the initial estimator is consistent for the target parameter, the consistency property is conserved through the targeting step. This additional step may even remove finite sample bias. Another desirable property is efficiency. When both the initial estimator and the estimator of the treatment mechanism are consistent, then the TMLE will also be asymptotically efficient. Since TMLEs incorporate machine learning procedures, we make every effort in practice to achieve minimal bias and the asymptotic semiparametric efficiency bound for the variance. The TMLE for (2) is also a substitution estimator, which means it takes an estimator of $Q_0$ and plugs it into $\Psi()$. Substitution estimators are more robust to data sparsity and outliers than nonsubstitution estimators.

The TMLE procedure can be summarized succinctly as follows: We take our initial estimator of $Q_0$ and then create a one-dimensional parametric statistical model with parameter $\varepsilon$ through the initial estimator whose score at $\varepsilon = 0$ spans the efficient influence curve of the parameter of interest at the initial estimator. $\varepsilon$ is estimated with maximum likelihood estimation in this parametric statistical model and we update the new estimator as the corresponding fluctuation of the initial estimator. The algorithm is iterated until convergence, although in our case for the risk difference, as in many others, it converges in one step. We use our celiac study with its simplified basic data structure and our risk difference target parameter to illustrate the TMLE. However, we note that the TMLE can estimate many varied target parameters.

### Implementation

**1. Estimating $\bar{Q}_0$.** In the previous section, we generated a super learner function for our celiac study. Therefore we used the super learner procedure to generate an initial estimate of $\bar{Q}_0$, denoted $\bar{Q}_n^0$.

**2. Estimating $g_0$.** Our targeting step required an estimate of the conditional distribution of diet compliance given covariates $W$. This estimate of $P_0(A \mid W) \equiv g_0$ is denoted $g_n$ and was obtained using super learning.

**3. Determining a parametric working model to fluctuate the initial estimator.** We use the estimate $g_n$ in a clever covariate to define a parametric working model coding fluctuations of the initial estimator. The clever covariate $H_n^*(A,W)$ is given by

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

**4. Updating $\bar{Q}_n^0$.** We ran a logistic regression of $Y$ on $H_n^*(A,W)$ using the offset $\text{logit}\bar{Q}_n^0(A,W)$ as the intercept in order to obtain $\varepsilon_n$, which is the estimated coefficient for $H_n^*(A,W)$. In order to update the estimate $\bar{Q}_n^0$ into a new (and final) estimate $\bar{Q}_n^1$ we calculated:

$$\text{logit}\,\bar{Q}_n^1(A,W) = \text{logit}\,\bar{Q}_n^0(A,W) + \varepsilon_n H_n^*(A,W).$$

This step would be iterated if necessary, however, for this TMLE it converges in one step.

The TMLE of $Q_0$ was therefore given by $Q_n^* = (\bar{Q}_n^1, Q_{W,n}^0)$.

**5. Targeted substitution estimator of the target parameter.** The plug-in TMLE using updated estimates $\bar{Q}_n^1(1,W)$ and $\bar{Q}_n^1(0,W)$ and the empirical distribution of $W$ was calculated. Our formula (3) became

$$\psi_{TMLE,n} = \Psi(Q_n^*) = \frac{1}{n}\sum_{i=1}^{n}\{\bar{Q}_n^1(1,W_i) - \bar{Q}_n^1(0,W_i)\}.$$

**6. Inference.** The influence curve of our estimator, defined as:

$$IC_n(O_i) = \left( \frac{I(A_i=1)}{g_n(1 \mid W_i)} - \frac{I(A_i=0)}{g_n(0 \mid W_i)} \right)(Y - \bar{Q}_n^1(A_i,W_i)) + \bar{Q}_n^1(1,W_i) - \bar{Q}_n^1(0,W_i) - \psi_{TMLE,n},$$

is used to calculate standard errors. With the influence curve, one can go forward with statistical inference as if the estimator minus its estimand equals the empirical mean of the influence curve. The sample mean of the estimated influence curve values was $\bar{IC}_n = \frac{1}{n}\sum_{i=1}^{n} IC_n(o_i)$. For our TMLE we have $\bar{IC}_n = 0$. We then calculated the sample variance of the estimated influence curve values:

$$S^2(IC_n) = \frac{1}{n}\sum_{i=1}^{n}\left(IC_n(o_i) - \bar{IC}_n\right)^2,$$

and the standard error of our estimator was given by $\sigma_n = \sqrt{S^2(IC_n)/n}$. With these standard errors, we calculated confidence intervals and $p$-values in the standard fashion. A 95% Wald-type confidence interval was given by $\psi_{TMLE,n} \pm z_{0.975}\sigma_n/\sqrt{n}$, where $z_\alpha$ denoted the $\alpha$-quantile of the standard normal density $N(0,1)$. The $p$-value for $\psi_{TMLE,n}$ was constructed as:

$$2\left[1 - \Phi\left(\left|\frac{\psi_{TMLE,n}}{\sigma_n/\sqrt{n}}\right|\right)\right],$$

where $\Phi$ denotes the standard normal cumulative distribution function.

**7. Interpretation.** The interpretation of our estimate $\psi_{TMLE,n}$, under causal assumptions, would be that noncompliance on a gluten-free diet causally increases/decreases (depending on the estimate and confidence interval) 10-year intestinal cancer risk in the United States celiac population. If our causal effect was not identifiable, perhaps due to unmeasured confounding, then we would interpret our estimate $\psi_{TMLE,n}$ purely statistically. For example, the additive effect of noncompliance on a gluten-free diet, controlling for measured confounders only, increases/decreases 10-year intestinal cancer risk in the United States celiac population by $\psi_{TMLE,n}$.

# Targeted Learning Framework

While we focused in this chapter on a simplified study of diet compliance on intestinal cancer in celiac disease subjects and a marginal risk difference parameter, targeted learning methodology is truly a framework for an algorithm that can be applied and developed for other study designs and data structures (van der Laan and Rubin 2006; van der Laan and Rose 2011). These problems can be attacked with the roadmap we have described in this chapter: 1) define the data, 2) state the model, 3) identify the target parameter, and 4) implement (and possibly develop a new) TMLE incorporating super learner providing inference with influence curves or bootstrapping.

One common issue found in practice is a right-censored outcome. In our fictional study of intestinal cancer, we are likely to have subjects drop-out, since follow-up occurs over a 10 year time frame. Drop-out may occur before the final 10 year endpoint due to death, relocation, disinterest in continuation, and other reasons. We can apply targeted learning to this data structure, which we write as $O = (W, A, \tilde{T}, \Delta) \sim P_0$, where we now additionally have $T$ for time to event $Y$, $C$ as censoring time, $\tilde{T} = \min(T, C)$, $\Delta = I(T \leq \tilde{T}) = I(C \geq T)$, with $Y$, $A$, and $W$ as defined previously. See, for example, Moore and van der Laan (2009a,c).

This leads us to another type of study design. In the study of rare diseases, such as intestinal cancer, researchers often use a so-called case-control design that samples diseased subjects from the celiac population of individuals with intestinal cancer, and then samples nondiseased subjects from the celiac population of individuals without intestinal cancer. This design eliminates the need to sample subjects and then follow them for extended periods. Case-control studies are biased designs, as the distribution of cases in the sample is not the same as in the target population. These designs also vary in complexity, and can include matching, incidence-density sampling, and other variations. Targeted learning methods can be applied with use of the known or estimated prevalence probability in case-control-weighted TMLE (van der Laan 2008; Rose and van der Laan 2008, 2009). A related design, often referred to as a nested case-control study, is a two-stage design, where one first samples a cohort from the population of interest, and then samples cases and controls from the cohort. An inverse probability weighted TMLE has been developed for use in these designs (Rose and van der Laan 2011).

Another design we highlight, is that of a community RCTs. This involves very carefully describing the observational unit, since interventions are assigned to communities and not individuals. These studies may also involve matching and follow subjects over time. Certain questions of interest are the causal effect of the community level intervention, and the direct effect of the community level intervention on the individual versus the indirect effect through other individuals

(van der Laan 2010).

Longitudinal studies are quite common, and often studies analyzed at two time points (baseline and endpoint) actually collect data at multiple time points, and this information is ignored. But what if we care about the effect of diet over these many time points? One cannot even define a coefficient in a parametric model that targets this effect. Suppose we observe $n$ i.i.d. copies of the longitudinal data structure $O = (L_0, A_0, \ldots, L_K, A_K, Y = L_{K+1})$, with $A_j$ a discrete valued intervention node, $L_j$ an intermediate covariate that occurs after $A_{j-1}$ and before $A_j$, $j = 0, \ldots, K$, and $Y$ is the final outcome of interest. For example, the TMLE for the intervention specific mean outcome will be a targeted plug-in estimator of: $E_{P^a} Y^a = E \ldots E(E(Y^a \mid \bar{L}_K^a) \mid L_{K-1}^a) \ldots \mid L_0)$, where $P^a$ is the post-intervention distribution setting $\bar{A}_K = (A_0, A_1, \ldots, A_K)$ equal to $\bar{a}_K = (a_0, a_1, \ldots, a_K)$ and $L^a = (L_0, L_1^a, \ldots, Y^a = L_{K+1}^a)$ is the random variable with probability distribution $P^a$. Other target parameters can also be estimated (van der Laan and Gruber 2012).

# Practical Implications

Many of the concepts presented in this chapter may be new and challenging, and the reader may wonder, "Is all this effort worth it? Why can't I just use standard methods as others have done for many decades?" The reality is that, all other arguments aside, conventional methods cannot accommodate the new era of "Big Data." They cannot effectively incorporate hundreds of covariates or multiple time point interventions. It has never been more important to accurately define your data, a realistic model, and target parameter, and then use flexible nonparametric (or semiparametric) methods to estimate your parameter of interest.

We motivated this chapter using the example of celiac disease. Experts studying this disease right now are exploring new areas, such as the timing of gluten introduction in young infants and their microbial environment, in order to understand why certain people with a genetic predisposition develop the disorder, but others with the genetic predisposition do not (Sellitto et al. 2012). These studies produce vast amounts of data, incredibly high-dimensional data sets, requiring novel flexible nonparametric statistical methods. If we specified a parametric statistical model in these studies, we would have more unknown parameters in the model than actual observations! Celiac disease is just one example. Brain imaging data is another area where researchers are inundated with vast amounts of data. Most of these data sets are comprised of groupings of 3 dimensional arrays over time, which leads to a 4 dimensional array for each individual in the study. Again, off-the-shelf parametric methods are not be able to handle the complexity of neuroimaging data. Comprehensive medical databases, such as the U.S. Food and Drug Administration's Sentinel Initiative, are another example of "Big Data." The Sentinel Initiative's goal is to monitor drugs and devices for safety over time. In order to do this, they are building a national electronic system which already has access to 100 million subjects. It is very unlikely that parametric modeling assumptions will be supported by subject matter knowledge, and there are many additional complications in data of this type: missingness, censoring, and time-dependent confounding are just a few (Rose 2012b).

Thus, the TMLE has increasingly important implications, both in the analysis of RCTs and observational studies. In RCTs, the TMLE often results in both bias and efficiency gains when compared to unadjusted and other ad hoc estimators frequently used in RCT data. This is due

to the fact that TMLE naturally exploits information in baseline and time-dependent covariates, reducing bias due to empirical confounding, as well as "drop out." We need not assume that this drop out is non informative, as in many cases it is informative (e.g., subjects experiencing toxicity side effects are more likely to drop out than those who do not).

Because we lack randomization in observational studies, efficient and maximally unbiased estimators are critical. Consider an observational study where people take a high dose or low dose of a medication. Further suppose the high dose is taken by sicker people. An unadjusted estimator and a misspecified parametric regression will be biased. A targeted learning approach gives us a robust and semiparametric efficient substitution estimator, while we maintain an appropriate loss function as the primary criterion.

# Some Philosophical and Methodological Implications

In the previous sections we pointed out that targeted maximum likelihood estimation and super learning framework (TMLE/SL) is not yet another set or collection of analytical tools to be added to the toolkit of the data-analyst, empirical research worker, or statistical consultant. Rather than this, it establishes a new methodology in several ways. From a technical point of view it offers *an integrative approach to data-analysis or statistical learning* by combining mathematical statistics with techniques from the field of computational intelligence (machine learning, data mining, knowledge discovery in databases). From a conceptual or methodological point of view, it sheds new lights on several "stages" of the research process, including such items as the research question, assumptions and background knowledge, modeling, causal inference, and validation, by anchoring these stages or elements of the research process in statistical theory. All these elements should be related to or defined in terms of (properties of) the data-generating distribution, and to this aim TMLE/SL provides both clear heuristics and formal underpinnings. Among other things, this means that the concept of a *statistical model* is reestablished in a parsimonious way, allowing humans to include only their *true*, realistic knowledge in the model. In addition, the scientific question and background knowledge are to be translated into a formal causal model and target causal parameter using the causal graphs and counterfactual (potential outcome) frameworks, including specifying a working marginal structural model. And, even more significantly, TMLE/SL reassigns to the very concept of *estimation*, canonical as it has always been in *statistical inference*, the pivotal role in any theory of/approach to "learning from data," whether it deals with establishing causal relations, classifying, clustering, multiple testing, or time series forecasting. This remark may seem rather obvious or even trivial from a statistical perspective, but appears to be crucial in understanding the situation in research practice today.

However, the relevance of TMLE/SL exceeds the realms of statistics in a strict sense and even those of research methodology. It has also several important philosophical implications, more specifically in the field of philosophy of science and epistemology. Some of these implications have been treated in some detail in Starmans (2011a) where the present state of statistical data analysis was put in a *historical and philosophical perspective* with the purpose *to clarify, understand*, and *account for* the current situation in statistical data analysis and to underline the relevance of topics addressed by TMLE for both philosophy of statistics and current issues in epistemology/philosophy of science. In this respect the current situation in data analysis is rather

paradoxical and inconvenient. From a foundational perspective the field consists of several competing schools with sometimes incompatible principles, approaches or viewpoints. Some can be traced back to Karl Pearson's approach to data-analysis, the Fisherian tradition of estimation, the Neyman–Pearson school of hypothesis testing, or the Bayesian paradigm. Simultaneously, the market for statistical textbooks offers many elementary and even advanced studies, raising the false impression of a uniform and united field with foundations that are fixed, and on which full agreement has been reached. It offers a toolkit that wrongly suggest the unification of ideas and methods derived from the aforementioned traditions. From a philosophical point of view this situation is rather inconvenient for two related reasons.

First, nearly all scientific disciplines have experienced a probabilistic revolution since the late 19th century. Increasingly, their key notions are probabilistic, their research methods, entire theories are probabilistic, if not the underlying worldview is probabilistic, i.e., dominated by and rooted in probability theory and statistics. When the probabilistic revolution emerged in the late 19th century, we observed this transition in renowned sciences like physics (kinetic gas theory, statistical mechanics), but especially in new emerging disciplines like the social sciences, biology (evolution, genetics, zoology), agricultural science, and psychology. Biology even came to maturity due to close interaction with statistics. Today, this trend has only further strengthened, and as a result there is a plethora of fields of application of statistics ranging from biostatistics, geostatistics, epidemiology, and econometrics to actuarial science, statistical finance, quality control and operational research in industrial engineering, and management science. Probabilistic approaches have also intruded several branches of computer science; most noticeably they dominate artificial intelligence.

Secondly, at a more abstract level, probabilistic approaches also dominate epistemology, the branch of philosophy committed to classical questions like: What is reality? Does it exist mind-independent? Do we have access to it? If yes, how? Can we make true statements about it? If yes, what is truth and how is it connected to reality? The analyses conducted to analyze these issues are usually intrinsically probabilistic. As a result these approaches dominate key-issues and controversies in epistemology, such as the scientific realism debate, the structure of scientific theories, Bayesian confirmation theory, causality, models of explanation and the status of natural laws. All too often scientific reasoning seems nearly synonymous with probabilistic reasoning. Considering the fact that scientific inference more and more depends on probabilistic reasoning and that statistical data analysis is not as well-founded as might be expected, the issue addressed in this chapter is of major importance for epistemology. If one views TMLE/SL against this background several philosophical implications become apparent. We confine ourselves to three salient issues: the erosion of models, the problem of causality, and the scope of the Bayesian paradigm.

Regarding the status of models it is obvious that they are ubiquitous in science and in everyday life. They are primarily commonsense notions or natural categories, that we constantly use to interpret reality, to understand ourselves, our situatedness and experiences. Despite many types of models and levels of abstraction the general idea is clear: they represent, describe or build up (a part or aspect of) reality for the purpose of interpreting, understanding or predicting this reality. In epistemology, the model theoretic tradition was adopted after successful application in mathematics. In this semantical tradition models are true interpretations of theories, and notions like logical consequence and validity are essentially semantical or model-based. As we pointed out in a previous section the notion of model in statistics fits this tradition. Models are essentially

sets of distributions including the approximation of the true data generating distribution. However, currently the concept of models seems to some extent eroded and there appears to be a tension between truth and models, both in epistemology and in research practice (Starmans 2011b). In epistemology, no doubt, pragmatic and postmodern approaches to science have contributed to the erosion of the concept of models. The most radical critic of the outlined representation might be Richard Rorty in his influential "Philosophy and the Mirror of Nature" (Rorty 1979). Emanating from the pragmatic tradition of Peirce, James, and Dewey and building on such diverse thinkers as Quine and Kuhn, he rejected the view that modern science is capable of an objective picture or representation of reality and contended that the postulated correspondence between language, mental image, and reality is a misconception. Even the famous transition from mythos to logos, which according to tradition, started about 2600 years ago, is unsustainable and the search for truth and the foundations of knowledge is, according to Rorty, illusory. But also in the sciences, models are often used instrumentally, making truth an obsolete and unnecessary notion. Although these ideas are also pervasive in the natural sciences, we confine ourselves here to statistics. This last idea culminates in particular in the work of the British statistician George Box, who in an infamous quote from 1987 states that "Essentially, all models are wrong, but some are useful" (Box and Draper 1987). With this passage, he refers to a research practice that even today is manifest in many sciences and in professional practice. Statistical data analysis usually takes place by means of (parametric) statistical models, which rely heavily on various assumptions about the population to which the analyzed data were derived. These assumptions are generally false, but the models are still widely used, supported by widely available and increasingly cheap computing power. By using parsimonious models, TMLE may contribute to the much needed reconciliation between the notions of models and truth.

Next, we refer to the issue of causality, which is principally a commonsense notion or natural category in the Aristotelian or Kantian sense of the word. Despite the major role of causality in both human action and in the history of philosophy and science, the notion proved problematic in the course of the history of ideas and sometimes even received a pejorative interpretation. Several attempts were made to ban the concept from scientific vocabulary, most noticeably by the philosopher and mathematician Bertrand Russell and leading 19th century statistician Karl Pearson. Thanks to fundamental research in computer science and artificial intelligence in the last decades, a computational turn took place in thinking about causality, which gave scientists increasingly a firm grip on this formerly unruly concept. In 2012, the American computer scientist and AI scientist Judea Pearl received the most prestigious scientific award in computer science, the Turing Award, for his work in the field of probabilistic networks and causal modeling. His 2000 publication and 2009 revised study "Causality: Models, Reasoning and Inference" (Pearl 2009) may be regarded as a highlight in modern thinking about causality. TMLE/SL uses the graphical methods pioneered by Pearl, but moves a step further and anchors it in statistical theory on estimation by relating the causal interpretation to the statistical model in the way we outlined in this chapter.

Thirdly, we briefly mention Bayesianism. Today, many philosophers of science to a greater or lesser extent embrace the Bayesian confirmation theory, in particular those who still pursue a rational reconstruction of science, after the failures or shortcomings of the inductive logic of the logical positivists and after the limitations of Karl Poppers falsificationism became apparent. The central question is how we put our faith in scientific theories or hypotheses, and especially how we can justify our faith in these theories on the basis of new empirical data. In statistics, computer

science, and Artificial Intelligence, Bayesianism appears as a leading paradigm. However, despite its intuitive appeal, the Bayesian approach suffers from the same drawbacks we outlined in this chapter, mainly because the approach is generally not targeted. Neither does it anchor the different "steps" of the research process, nor does it provide the empirical researcher with a constructive methodology. In view of all these aspects it is certainly not fully evident to identify Bayesianism as the canonical model for scientific knowledge and reasoning.

# Notes and Further Reading

Targeted learning methods are further expanded in van der Laan and Rose (2011), and portions of the introductory material in this chapter were adapted from Chapters 1-5 of that textbook as well as Starmans (2011a). Detailed references concerning related work can also be found in those chapters, as this section is not exhaustive. The general TMLE algorithm was published in van der Laan and Rubin (2006), and additional papers on TMLE include Bembom and van der Laan (2007), van der Laan (2008), Rose and van der Laan (2008, 2009, 2011), Moore and van der Laan (2009a,b,c), Bembom et al. (2009), Polley and van der Laan (2009), Rosenblum et al. (2009), van der Laan and Gruber (2010, 2012), Gruber and van der Laan (2010), Rosenblum and van der Laan (2010), and Wang et al. (2010).

The super learner was published in van der Laan et al. (2007). An introductory implementation article using an application in an elderly population can be found in Rose (2012a). There is significant related work on ensembling that predated the original super learner paper, and we refer to Chapter 3 of van der Laan and Rose (2011) for more complete references. We do note here that the super learner is a generalization of the stacking algorithm (Wolpert 1992; Breiman 1996; LeBlanc and Tibshirani 1996) and the name "super learner" was coined due to the theoretical oracle property and its ramifications described in van der Laan and Dudoit (2003).

There are other estimators in the literature that can estimate our described parameter. The original maximum-likelihood-based substitution estimator of the g-formula parameter was published in Robins (1986). An introductory implementation of this estimator was recently published (Snowden et al. 2011; Rose et al. 2011). A thorough bibliographic review of estimating equation methods, which includes inverse probability weighting (Robins 1999; Hernan et al. 2000) and augmented inverse probability weighting (Robins et al. 2000b; Robins 2000; Robins and Rotnitzky 2001), is covered in Chapter 1 of van der Laan and Robins (2003). Introductory implementations of inverse probability weighting are described in Robins et al. (2000a), Mortimer et al. (2005), and Cole and Hernan (2008). These estimators lack the complete statistical properties of TMLE, as described in Chapter 6 of van der Laan and Rose (2011).

The SCM framework discussed in this chapter is given a thorough treatment in Pearl (2009). An alternative causal framework is the Neyman–Rubin causal model (Neyman 1923; Rubin 1974). We motivated this chapter using an example from celiac disease research. Celiac disease is a common autoimmune disorder, and currently the only autoimmune disorder with a known treatment: the gluten-free diet. Pertinent references include Green et al. (2001) and Fasano et al. (2003).

## Acknowledgements

SR was supported by a fellowship from the National Science Foundation (Grant DMS-1103901). MvdL was supported by the National Institutes of Health (Grant R01 A1074345-01).

## References

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

O. Bembom, M.L. Petersen, S.-Y. Rhee, W.J. Fessel, S.E. Sinisi, R.W. Shafer, and M.J. van der Laan. Biomarker discovery using targeted maximum likelihood estimation: application to the treatment of antiretroviral resistant HIV infection. *Stat Med*, 28:152–72, 2009.

G.E.P. Box and N.R. Draper. *Empirical Model-Building and Response Surfaces*. Wiley, Hoboken, 1987.

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

S.R. Cole and M.A. Hernan. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*, 168:656–664, 2008.

A. Fasano, I. Berti, T. Gerarduzzi, T. Not, R. Colletti, S. Drago, Y. Elitsur, P. Green, S. Guandalini, I. Hill, M. Pietzak, A. Ventura, M. Thorpe, D. Kryszak, F. Fornaroli, S. Wasserman, J. Murray, and K. Horvath. Prevalence of celiac disease in at-risk and not-at-risk groups in the United States: A large multicenter study. *Arch Intern Med*, 163(3):28692, 2003.

P. Green, S. Stravropoulos, S. Pangagi, S. Goldstein, D. McMahon, H. Absan, and A. Neugut. Characteristics of adult celiac disease in the USA: Results of a national survey. *Am J Gastroenterol*, 96:12631, 2001.

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

M.A. Hernan, B. Brumback, and J.M. Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiol*, 11(5):561–570, 2000.

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

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

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

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

K.M. Mortimer, R. Neugebauer, M.J. van der Laan, and I.B. Tager. An application of model-fitting procedures for marginal structural models. *Am J Epidemiol*, 162(4):382–388, 2005.

J. Neyman. Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes (In Polish). English translation by D.M. Dabrowska and T.P. Speed (1990). *Stat Sci*, 5:465–480, 1923.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge, New York, 2nd edition, 2009.

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

J.M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods–application to control of the healthy worker survivor effect. *Math Mod*, 7:1393–1512, 1986.

J.M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology: The Environment and Clinical Trials*. Springer, Berlin Heidelberg New York, 1999.

J.M. Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, 2000.

J.M. Robins and A. Rotnitzky. Comment on the Bickel and Kwon article, "Inference for semiparametric models: some questions and an answer". *Stat Sinica*, 11(4):920–936, 2001.

J.M. Robins, M.A. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiol*, 11(5):550–560, 2000a.

J.M. Robins, A. Rotnitzky, and M.J. van der Laan. Comment on "On profile likelihood". *J Am Stat Assoc*, 450:431–435, 2000b.

R. Rorty. *Philosophy and the Mirror of Nature*. Princeton University Press, Princeton, 1979.

S. Rose. Mortality risk score prediction in an elderly population using machine learning. *Am J Epidemiol*, in press, 2012a.

S. Rose. Big data and the future. *Significance*, 9(4):4748, 2012b.

S. Rose and M.J. van der Laan. Simple optimal weighting of cases and controls in case-control studies. *Int J Biostat*, 4(1):Article 19, 2008.

S. Rose and M.J. van der Laan. Why match? Investigating matched case-control study designs with causal effect estimation. *Int J Biostat*, 5(1):Article 1, 2009.

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

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

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

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

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*, 66:688–701, 1974.

M. Sellitto, G. Bai, G. Serena, W. Fricke, C. Sturgeon, P. Gajer, J. White, S. Koenig, J. Sakamoto, D. Boothe, R. Gicquelais, D. Kryszak, E. Puppa, C. Catassi, J. Ravel, and A. Fasano. Proof of concept of microbiome-metabolome analysis and delayed gluten exposure on celiac disease autoimmunity in genetically at-risk infants. *PLoS ONE*, 7(3):e33387, 2012. doi: 10.1371/journal.pone.0033387.

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

R.J.C.M. Starmans. Models, inference and truth: Probabilistic reasoning in the information era. In M.J. van der Laan and S. Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, Berlin Heidelberg New York, 2011a.

R.J.C.M. Starmans. The reality behind the model and the cracks in the mirror of nature (in Dutch). *Filosofie*, 21(5):3943, 2011b.

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

M.J. van der Laan. Estimation of causal effects of community-based interventions. Technical Report 268, Division of Biostatistics, University of California, Berkeley, 2010.

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

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

M.J. van der Laan and S. Gruber. Targeted minimum loss based estimation of causal effects of multiple time point interventions. *Int J Biostat*, 8(1), 2012.

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

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

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

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

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

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