Entering the Era of Data Science: Targeted Learning and the Integration of Statistics and Computational Data Analysis

Mark J. van der Laan[*] Richard J.C.M. Starmans[†]

[*] University of California, Berkeley, Division of Biostatistics, laan@berkeley.edu

[†] Universiteit Utrecht, Department of Information and Computing Sciences, R.J.C.M.Starmans@uu.nl

Entering the Era of Data Science: Targeted Learning and the Integration of Statistics and Computational Data Analysis

Mark J. van der Laan and Richard J.C.M. Starmans

Abstract

This outlook article will appear in Advances in Statistics and it reviews the research of Dr. van der Laan's group on Targeted Learning, a subfield of statistics that is concerned with the construction of data adaptive estimators of user-supplied target parameters of the probability distribution of the data and corresponding confidence intervals, aiming to only rely on realistic statistical assumptions. Targeted Learning fully utilizes the state of the art in machine learning tools, while still preserving the important identity of statistics as a field that is concerned with both accurate estimation of the true target parameter value and assessment of uncertainty in order to make sound statistical conclusions. We also provide a philosophical historical perspective on Targeted Learning, also relating it to the new developments in Big Data. We conclude with some remarks explaining the immediate relevance of Targeted Learning to the current big data movement.

1 Introduction

In Section 2 we start out with reviewing some basic statistical concepts such as data probability distribution, statistical model, and target parameter, allowing us to define the field Targeted Learning, a sub-field of statistics that develops data adaptive estimators of user supplied target parameters of data distributions based on high dimensional data under realistic assumptions (e.g., incorporating the state of the art in machine learning) while preserving statistical inference. This also allows us to clarify how Targeted Learning distinguishes from typical current practice in data analysis that relies on unrealistic assumptions, and describe the key ingredients of targeted minimum loss based estimation (TMLE), a general tool to achieve the goals set out by Targeted Learning: a substitution estimator, construction of initial estimator through super-learning, targeting of the initial estimator to achieve asymptotic linearity with known influence curve by solving the efficient influence curve estimating equation, and statistical inference in terms of a normal limiting distribution.

Targeted learning resurrects the pillars of statistics such as that a model represents actual knowledge about the data generating experiment and that a target parameter represents the feature of the data generating distribution we want to learn from the data. In this manner, targeted learning defines a truth, and sets a scientific standard for estimation procedures, while current practice typically defines a parameter as a coefficient in a misspecified parametric model (e.g., logistic linear regression, repeated measures generalized linear regression) or small unrealistic semi parametric regression models (e.g., Cox proportional hazards regression), where different choices of such misspecified models yield different answers. This lack of truth in current practice, supported by statements such as ”All models are wrong but some are useful”, allows a user to make arbitrary choices even though these choices result in different answers to the same estimation problem. In fact, this lack of truth in current practice presents a fundamental drive behind the epidemic of false positives and lack of power to detect true positives our field is suffering from. In addition, this lack of truth makes many of us question the scientific integrity of the field we call statistics and makes it impossible to teach statistics as a scientific discipline, even though the foundations of statistics, including a very rich theory, are purely scientific. That is, our field has suffered from a disconnect between the theory of statistics and the practice of statistics, while practice should be driver by relevant theory and theoretical developments should be driven by practice. For example, a theorem establishing consistency and asymptotic normality of a maximum likelihood estimator for a parametric model that is known to be misspecified is not a relevant theorem for practice since the true data generating distribution is not captured by this theorem.

Defining the statistical model to actually contain the true probability distribution has enormous implications for the development of valid estimators. For example, maximum likelihood estimators are now ill defined due to the curse of dimensionality of the model. In addition, even regularized maximum likelihood estimators are seriously flawed: A general problem with maximum likelihood based estimators is that the estimate the density of the data distribution, where the maximum likelihood criterion only cares about how well the density estimators fits the true density, resulting in a wrong trade-off for the actual target parameter of interest. From a practical perspective, when we use AIC, BIC, or cross-validated log-likelihood to select variables in our regression model, then that procedure is ignorant of the specific feature of the data distribution we want to estimate. That is, in large statistical models it is immediately apparent that estimators need to be targeted towards their goal, just like a human being learns the answer to a specific question in a targeted manner, and maximum likelihood based estimators fail to do that.

In Section 3 we review the roadmap for targeted learning of a causal quantity, involving defining a causal model and causal quantity of interest, establishing an estimand of the data distribution that equals the desired causal quantity under additional causal assumptions, applying the pure statistical targeted learning of the relevant estimand based on a statistical model compatible with the causal model but for sure containing the true data distribution, and careful interpretation of the results. In Section 4 we proceed with describing our proposed targeted minimum loss-based estimation (TMLE) template, which represents a concrete template for construction of targeted efficient substitution estimators which are not only asymptotically consistent, asymptotically normally distributed, and asymptotically efficient, but are also tailored to have robust finite sample performance. Subsequently, in Section 5 we review some of our most important advances in Targeted Learning, demonstrating the remarkable power and flexibility of this TMLE methodology, and in Section 6 we describe future challenges and areas of research. In Section 7 we provide a historical philosophical perspective of Targeted Learning. Finally, in Section 8 we conclude with some remarks, putting Targeted Learning in the context of the modern era of Big Data.

We refer to our papers and book on targeted learning for overviews of relevant parts of the literature that put our specific contributions within the field of Targeted Learning in the context of the current literature, thereby allowing us to focus on targeted learning itself in the current outlook article.

2 Targeted Learning

Our research takes place in a sub-field of statistics we named Targeted Learning (82; 47). In statistics the data $(O_1, \ldots, O_n)$ on n units is viewed as a realization of a random variable, or equivalently, an outcome of a particular experiment, and thereby has a probability distribution P_0^n, often called the data distribution. For example, one might observe $O_i = (W_i, A_i, Y_i)$ on a subject i, where W_i are baseline characteristics of the subject, A_i is a binary treatment or exposure the subject received, and Y_i is a binary outcome of interest such as an indicator of death, $i = 1, \ldots, n$. Throughout this article we will use this data structure to demonstrate the concepts and estimation procedures.

2.1 Statistical model

A statistical model $\mathcal{M}^n$ is defined as a set of possible probability distributions for the data distribution, and thus represents the available statistical knowledge about the true data distribution P_0^n. In Targeted Learning, this core-definition of the statistical model is fully respected in the sense that one should define the statistical model to contain the true data distribution: $P_0^n \in \mathcal{M}^n$. So contrary to the often conveniently used slogan "All models are wrong, but some are useful" and erosion over time of the original true meaning of a statistical model throughout applied research, Targeted Learning defines the model for what it actually is (59). If there is truly no statistical knowledge available, then the statistical model is defined as all data distributions. A possible statistical model is the model that assumes that $(O_1, \ldots, O_n)$ are n independent and identically distributed random variables with completely unknown probability distribution P_0, representing the case that the sampling of the data involved repeating the same experiment independently. In our example, this would mean that we assume that (W_i, A_i, Y_i) are independent with a completely unspecified common probability distribution. For example, if W is 10 dimensional, while A and Y are two dimensional, then P_0 is described by a 12-dimensional density, and this statistical model does not put any restrictions on this 12-dimensional density. One could factorize this density of (W, A, Y) as follows:

$$p_0(W, A, Y) = p_{W,0}(W) p_{A|W,0}(A \mid W) p_{Y|A,W,0}(Y \mid A, W),$$

where $p_{W,0}$ is the density of the marginal distribution of W, $p_{A|W,0}$ is the conditional density of A, given W, and $p_{Y|A,W,0}$ is the conditional density of Y, given A, W. In this model, each of these factors are unrestricted. On the other

other hand, suppose now that the data is generated by a randomized controlled trial in which we randomly assign treatment $A \in \{0, 1\}$ with probability 0.5 to a subject. In that case, the conditional density of A, given W, is known, but the marginal distribution of the covariates and the conditional distribution of the outcome, given covariates and treatment, might still be unrestricted. Even in an observational study, one might know that treatment decisions where only based on a small subset of the available covariates W, so that it is known that $p_{A|W,0}(1 \mid W)$ only depends on W through these few covariates. In the case that death $Y = 1$ represents a rare event, it might also be known that the probability of death $P_{Y|A,W}(1 \mid A, W)$ is known to be between 0 and some small number (e.g., 0.03). This restriction should then be included in the model $\mathcal{M}$.

In various applications, careful understanding of the experiment that generated the data might show that even these rather large statistical models assuming the the data generating experiment equals the independent repetition of a common experiment is too small to be true: see (69; 11; 10; 74; 71) for models in which $(O_1, \ldots, O_n)$ is a joint random variable described by a single experiment, which nonetheless involves a variety of conditional independence assumptions. That is, the typical statement that $O_1, \ldots, O_n$ are independent and identically distributed (i.i.d.) might already represent a wrong statistical model. For example, in a community randomized trial it is often the case that the treatments are assigned by the following type of algorithm: based on the characteristics $(W_1, \ldots, W_n)$, one first applies an algorithm that aims to split the n communities in $n/2$ pairs that are similar w.r.t. baseline characteristics, subsequently, one randomly assigns treatment and control to each pair. Clearly, even when the communities would have been randomly sampled from a target population of communities, the treatment assignment mechanism creates dependence so that the data generating experiment cannot be described as an independent repetition of experiments: see (74) for a detailed presentation.

In a study in which one observes a single community of n interconnected individuals one might have that the outcome Y_i for subject i is not only affected by the subject's past (W_i, A_i), but also affected by the covariate and treatment of friends of subject i. Knowing the friends of each subject i would now impose strong conditional independence assumptions on the density of the data $(O_1, \ldots, O_n)$, but one cannot assume that the data is a result of n independent experiments: in fact, as in the community randomized trial example, such data sets have sample size 1 since the data can only be described as the result of a single experiment (71).

In group sequential randomized trials, one often may use a randomization probability for a next recruited i-th subject that depends on the observed data of the previously recruited and observed subjects $O_1, \ldots, O_{i-1}$, which makes the treatment assignment A_i a function of $O_1, \ldots, O_{i-1}$. Even when the subjects are sampled randomly from a target population, this type of dependence between treatment A_i and the past data $O_1, \ldots, O_{i-1}$ implies that the data is the result of a single large experiment (again, the sample size equals 1) (69; 11; 10).

Indeed, many realistic statistical models only involve independence and conditional independence assumptions, and known bounds (e.g., it is known that the observed clinical outcome is bounded between [0,1] or the conditional probability of death is bounded between 0 and a small number). Either way, if the data distribution is described by a sequence of independent (and possibly identical) experiments or by a single experiment satisfying a variety of conditional independence restrictions, parametric models, although representing common practice, are practically always invalid statistical models since such knowledge about the data distribution is essentially never available.

An important by-product of requiring that the statistical model needs to be truthful is that one is forced to obtain as much knowledge about the experiment before committing to a model, which is precisely the role a good statistician should play. On the other hand, if one commits to a parametric model, then why would one still bother trying to find out the truth about the data generating experiment?

2.2 Target Parameter

The target parameter is defined as a mapping $\Psi : \mathcal{M}^n \rightarrow \mathbb{R}^d$ that maps the data distribution into the desired finite dimensional feature of the data distribution one wants to learn from the data: $\psi_0^n = \Psi(P_0^n)$. This choice of target parameter requires careful thought independent from the choice of statistical model, and is not a choice made out of convenience. The use of parametric or semi-parametric models such as the Cox-proportional hazards model are often accompanied with the implicit statement that the unknown coefficients represent the parameter of interest. Even in the unrealistic scenario that these small statistical models would be true, there is absolutely no reason why the very parametrization of the data distribution should correspond with the target parameter of interest. Instead, the statistical model $\mathcal{M}^n$ and the choice of target parameter $\Psi : \mathcal{M}^n \rightarrow \mathbb{R}^d$ are two completely separate choices, and, by no means, one should imply the other. That is, the statistical knowledge about the experiment that generated the data and defining what we hope to learn from the data are two important key steps in science that should not be convoluted. The true target parameter value ψ_0^n is obtained by applying the target parameter mapping Ψ to the true data distribution P_0^n, and represents the estimand of interest.

For example, if $O_i = (W_i, A_i, Y_i)$ are independent and have common probability distribution P_0, then one might define the target parameter as an average of the conditional W-specific treatment effects:

$$\psi_0 = \Psi(P_0) = E_{P_0}\{E_{P_0}(Y \mid A = 1, W) - E_{P_0}(Y \mid A = 0, W)\}.$$

By using that Y is binary, this can also be written as follows:

$$\psi_0 = \int_w \{P_{Y|A,W,0}(1 \mid A = 1, W = w) - P_{Y|A,W,0}(1 \mid A = 0, W = w)\}P_{W,0}(dw),$$

where $P_{Y|A,W,0}(1 \mid A = a, W = w)$ denotes the true conditional probability of death, given treatment $A = a$ and covariate $W = w$.

For example, suppose that the true conditional probability of death is given by some logistic function:

$$P_{Y|A,W}(1 \mid A, W) = \frac{1}{1 + \exp(-m_0(A, W))}$$

for some function m_0 of treatment A and W. The reader can plug-in a possible form for m_0 such as $m_0(a, w) = 0.3a + 0.2w_1 + 0.1w_1w_2 + aw_1w_2w_3$. Given this function m_0, the true value ψ_0 is computed by the above formula as follows:

$$\psi_0 = \int_w \left(\frac{1}{1 + \exp(-m_0(1, w))} - \frac{1}{1 + \exp(-m_0(0, w))}\right) P_{W,0}(dw).$$

This parameter ψ_0 has a clear statistical interpretaion as the average of all the w-specific additive treatment effects $\{P_{Y|A,W,0}(1 \mid A = 1, W = w) - P_{Y|A,W,0}(1 \mid A = 0, W = w)\}$.

2.3 The important role of models also involving non-testable assumptions

However, this particular statistical estimand ψ_0 has an even richer interpretation if one is willing to make additional so called causal (non-testable) assumptions. Let's assume that W, A, Y are generated by a set of so called structural equations:

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

where $U = (U_W, U_A, U_Y)$ are random inputs following a particular unknown probability distribution, while the functions f_W, f_A, f_Y deterministically map the realization of the random input $U = u$ sequentially into a realization of $W = f_W(u_W), A = f_A(W, u_A), Y = f_Y(W, A, u_y)$. One might not make any assumptions about the form of these functions f_W, f_A, f_Y. In that case, these causal assumptions put no restrictions on the probability distribution of (W, A, Y), but through these assumptions we have parametrized P_0 by a choice of functions (f_W, f_A, f_Y) and a choice of distribution of U. Pearl (35) refers to such assumptions as a structural causal model for the distribution of (W, A, Y).

This structural causal model allows one to define a corresponding post-intervention probability distribution that corresponds with replacing $A = f_A(W, U_A)$ by our desired intervention on the intervention node A. For example, a static intervention $A = 1$ results in a new system of equations $W = f_W(U_W), A = 1, Y_1 = f_Y(W, 1, U_Y)$, where this new random variable Y_1 is called a counterfactual outcome or potential outcome corresponding with intervention $A = 1$. Similarly, one can define $Y_0 = f_Y(W, 0, U_Y)$. Thus, Y_0 (Y_1) represent the outcome on the subject one would have seen if the subject would have been assigned treatment $A = 0$ $(A = 1)$. One might now define the causal effect of interest as $E_0Y_1 - E_0Y_0$, i.e. the difference between the expected outcome of Y_1 and the expected outcome of Y_0. If one also assumes that A is independent of U_Y, given W, which is often referred to as the assumption of no unmeasured confounding or the randomization assumption, then it follows that $\psi_0 = E_0Y_1 - E_0Y_0$. That is, under the structural causal model, including this no unmeasured confounding assumption, ψ_0 cannot only be interpreted purely statistically as an average of conditional treatment effects, but it actually equals the marginal additive causal effect.

In general, causal models or, more generally, sets of non-testable assumptions, can be used to define underlying target quantities of interest, and corresponding statistical target parameters that equal this target quantity under these assumptions. Well known classes of such models are models for censored data in which the observed data is represented as a many to one mapping on the full data of interest and censoring variable, and the target quantity is a parameter of the full data distribution. Similarly, causal inference models represent the observed data as a mapping on counterfactuals and the observed treatment (either explicitly as in the Neyman-Rubin model or implicitly as in the Pearl structural causal models), and one defines the target quantity as a parameter of the distribution of the counterfactuals. One is now often concerned with providing sets of assumptions on the underlying distribution (i.e., of the full-data) that allow identifiability of the target quantity from the observed data distribution (e.g. coarsening at random or randomization assumption). These non testable assumptions do not change the statistical model $\mathcal{M}$ and, as a consequence, once one has defined the relevant estimand ψ_0, do not affect the estimation problem either.

2.4 Estimation problem

The estimation problem is defined by the statistical model (i.e., $(O_1, \ldots, O_n) \sim P_0^n \in \mathcal{M}^n$), and choice of target parameter (i.e., $\Psi : \mathcal{M}^n \to \mathbb{R}$). Targeted learning is now the field concerned with the development of estimators of the target parameter that are asymptotical consistent as the number of units n converges to infinity, and whose appropriately standardized version (e.g., $\sqrt{n}(\psi_n - \psi_0^n)$) converges in probability distribution to some limit probability distribution (e.g., normal distribution), so that one can construct confidence intervals that for large enough sample size n contain with a user supplied high probability the true value of the target parameter. In the case that $O_1, \ldots, O_n \sim_{iid} P_0$, a common method for establishing asymptotic normality of an estimator is to demonstrate that the estimator minus truth can be approximated by an empirical mean of a function of O_i. Such an estimator is called asymptotically linear at P_0. Formally, an estimator ψ_n is asymptotically linear under i.id.. sampling from P_0 if $\psi_n - \psi_0 = \frac{1}{n}\sum_{i=1}^n IC(P_0)(O_i) + o_P(1/\sqrt{n})$, where $O \to IC(P_0)(O)$ is the so called influence curve at P_0. In that case, the central limit theorem teaches us that $\sqrt{n}(\psi_n - \psi_0)$ converges to a normal distribution $N(0, \sigma^2)$ with variance $\sigma^2 = E_{P_0} IC(P_0)(O)^2$ defined as the variance of the influence curve. An asymptotic 0.95-confidence interval for ψ_0 is then given by $\psi_n \pm 1.96\sigma_n/\sqrt{n}$, where σ_n^2 is the sample variance of an estimate $IC_n(O_i)$ of the true influence curve $IC(P_0)(O_i)$, $i = 1, \ldots, n$.

The empirical mean of the influence curve $IC(P_0)$ of an estimator ψ_n represents the first order linear approximation of the estimator as a functional of the empirical distribution, and the derivation of the influence curve is a by-product of the application of the so called functional delta-method for statistical inference based on functionals of the empirical distribution (18; 83; 19). That is, the influence curve $IC(P_0)(O)$ of an estimator, viewed as a mapping from the empirical distribution P_n into the estimated value $\hat{\Psi}(P_n)$, is defined as the directional derivative at P_0 in the direction $(P_{n=1} - P_0)$, where $P_{n=1}$ is the empirical distribution at a single observation O.

2.5 Targeted Learning respects both local and global constraints of the statistical model

Targeted learning is not just satisfied with asymptotic performance such as asymptotic efficiency. Asymptotic efficiency requires fully respecting the local statistical constraints for shrinking neighborhoods around the true data distribution implied by the statistical model, defined by the so called tangent space generated by all scores of parametric submodels through P_0^n (5), but it does *not* require respecting the global constraints on the data distribution implied by the statistical model (e.g., see (21)). Instead Targeted Learning pursues the development of such asymptotically efficient estimators that *also* have excellent and robust practical performance by also fully respecting the *global constraints* of the statistical model. In addition, Targeted Learning is also concerned with the development of confidence intervals with good practical coverage. For that purpose, our proposed methodology for Targeted Learning, so called targeted minimum loss based estimation discussed below, does not only result in asymptotically efficient estimators, but the estimators 1) utilize unified cross-validation to make practically sound choices for estimator construction that actually work well with the very data set at hand (75; 84; 76; 80; 38), 2) focusses on the construction of substitution estimators that by definition also fully respect the global constraints of the statistical model, and 3) uses influence curve theory to construct targeted computer friendly estimators of the asymptotic distribution, such as the normal limit distribution based on an estimator of the asymptotic variance of the estimator.

Let's succinctly review the immediate relevance to Targeted Learning of the above mentioned basic concepts: influence curve, efficient influence curve, substitution estimator, cross-validation and super-learning. For the sake of discussion, let's consider the case that the n observations are independent and identically distributed: $O_i \sim_{i.i.d.} P_0 \in \mathcal{M}$, and $\Psi : \mathcal{M} \rightarrow \mathbb{R}^d$ can now be defined as a parameter on the common distribution of O_i, but each of the concepts have a generalization to dependent data as well (e.g., see (71)).

2.6 Targeted Learning is based on a substitution estimator

Substitution estimators are estimators that can be described as the target parameter mapping applied to an estimator of the data distribution that is an element of the statistical model. More generally, if the target parameter is represented as a mapping on a part $Q_0 = Q(P_0)$ of the data distribution P_0 (e.g. factor of likelihood), then a substitution estimator can be represented as $\Psi(Q_n)$, where Q_n is an estimator of Q_0 that is contained in the parameter space $\{Q(P) : P \in \mathcal{M}\}$ implied by the statistical model $\mathcal{M}$. Substitution estimators are known to be particularly robust by fully respecting that the true target parameter is obtained by evaluating the target parameter mapping on this statistical model. For example, substitution estimators are guaranteed to respect known bounds on the target parameter (e.g. it is a probability or difference between two probabilities) as well as known bounds on the data distribution implied by the model $\mathcal{M}$.

In our running example, we can define $Q_0 = (Q_{W,0}, \bar{Q}_0)$, where $Q_{W,0}$ is the probability distribution of W under P_0, and $\bar{Q}_0(A, W) = E_{P_0}(Y \mid A, W)$ is the conditional mean of the outcome, given the treatment and covariates, and represent the target parameter

$$\psi_0 = \Psi(Q_0) = E_{Q_{W,0}}\{\bar{Q}_0(1, W) - \bar{Q}_0(0, W)\}$$

as a function of the conditional mean $\bar{Q}_0$ and the probability distribution $Q_{W,0}$ of W. The model $\mathcal{M}$ might restrict $\bar{Q}_0$ to be between 0 and a small number $delta < 1$, but otherwise puts no restrictions on Q_0. A substitution estimator is now obtained by plugging in the empirical distribution $Q_{W,n}$ for $Q_{W,0}$ and a data adaptive estimator $0 < \bar{Q}_n < \delta$ of the regression $\bar{Q}_0$:

$$\psi_n = \Psi(Q_{W,n}, \bar{Q}_n) = 1/n \sum_{i=1}^{n} \{\bar{Q}_n(1, W_i) - \bar{Q}_n(0, W_i)\}.$$

Not every type of estimator is a substitution estimator. For example, an inverse probability of treatment type estimator of ψ_0 could be defined as

$$\psi_{n,IPTW} = \frac{1}{n} \sum_{i=1}^{n} \frac{2A_i - 1}{G_n(A_i \mid W_i)} Y_i,$$

where $G_n(\cdot \mid W)$ is an estimator of the conditional probability of treatment $G_0(\cdot \mid W)$. This is clearly not a substitution estimator. In particular, if $G_n(A_i \mid W_i)$ is very small for some observations, this estimator might not be between -1 and 1, and thus completely ignores known constraints.

2.7 Targeted estimator relies on data adaptive estimator of nuisance parameter

The construction of targeted estimators of the target parameter requires construction of an estimator of infinite dimensional nuisance parameters, specifically the initial estimator of the relevant part Q_0 of the data distribution in the TMLE, and the estimator of the nuisance parameter $G_0 = G(P_0)$ that is needed to target the fit of this relevant part in the TMLE. In our running example, we have $Q_0 = (Q_{W,0}, \bar{Q}_0)$ and the G_0 is the conditional distribution of A, given W.

2.8 Targeted Learning uses super-learning to estimate the nuisance parameter

In order to optimize these estimators of the nuisance parameters (Q_0, G_0), we use a so called super-learner that is guaranteed to asymptotically outperform any available procedure by simply including it in the library of estimators that is used to define the super-learner.

The super-learner is defined by a library of estimators of the nuisance parameter, and uses cross-validation to select the best weighted combination of these estimators. The asymptotic optimality of the super-learner is implied by the oracle inequality for the cross-validation selector that compares the performance of the estimator that minimizes the cross-validated risk over all possible candidate estimators with the oracle selector that simply selects the best possible choice (as if one has available an infinite validation sample). The only assumption this asymptotic optimality relies upon is that the loss function used in cross-validation is uniformly bounded, and that the number of algorithms in the library does not increase at a faster rate than a polynomial power in sample size when sample size converges to infinity (75; 84; 76; 80; 38) However, cross-validation is a method that goes beyond optimal asymptotic performance, since the cross-validated risk measures the performance of the estimator on the very sample it is based upon, making it a practically very appealing method for estimator selection.

In our running example, we have that $\bar{Q}_0 = \arg\min_{\bar{Q}} E_{P_0} L(\bar{Q})(O)$, where $L(\bar{Q}) = (Y - \bar{Q}(A,W))^2$ is the squared error loss, or, one can also use the log-likelihood loss $L(\bar{Q})(O) = -\{Y \log \bar{Q}(A,W) + (1-Y)\log(1-\bar{Q}(A,W))\}$. Usually, there are a variety of possible loss functions one could use to define the super-learner: the choice could be based on the dissimilarity implied by the loss function (75), but probably should itself be data adaptively selected in a targeted manner. The cross-validated risk of a candidate estimator of $\bar{Q}_0$ is then defined as the empirical mean over a validation sample of the loss of the candidate estimator fitted on the training sample, averaged across different spits of the sample in a validation and training sample. A typical way to obtain such sample splits is so called V-fold cross-validation in which one first partitions the sample in V subsets of equal size, and each of the V subsets play the role of a validation sample while its complement of $V-1$ subsets equals the corresponding training sample. Thus, V-fold cross-validation results in V sample splits into a validation sample and corresponding training sample. A possible candidate estimator is a maximum likelihood estimator based on a logistic linear regression working model for $P(Y=1 \mid A, W)$. Different choices of such logistic linear regression working models result in different possible candidate estimators. So in this manner one can already generate a rich library of candidate estimators. However, the statistics and machine learning literature has also generated lots of data adaptive estimators based on smoothing, data adaptive selection of basis functions, and so on, resulting in another large collection of possible candidate estimators that can be added to the library. Given a library of candidate estimators, the super-learner selects the estimator that minimizes the cross-validated risk over all the candidate estimators. This selected estimator is now applied to the whole sample to give our final estimate $\bar{Q}_n$ of $\bar{Q}_0$. One can enrich the collection of candidate estimators by taking any weighted combination of an initial library of candidate estimators, thereby generating a whole parametric family of candidate estimators.

Similarly, one can define a super-learner of the conditional distribution of A, given W.

The super-learner's performance improves by enlarging the library. Even though for a given data set, one of the candidate estimators will do as well as the super-learner, across a variety of data sets, the super-learner beats an estimator that is betting on particular subsets of the parameter space containing the truth or allowing good approximations of the truth. The use of super-learner provides on important step in creating a robust estimator whose performance is not relying on being lucky but on generating a rich library so that a weighted combination of the estimators provides a good approximation of the truth, wherever the truth might be located in the parameter space.

2.9 Asymptotic efficiency

An asymptotically efficient estimator of the target parameter is an estimator that can be represented as the target parameter value plus an empirical mean of a so called (mean zero) efficient influence curve $D^*(P_0)(O)$, up till a second order term that is asymptotically negligible (5). That is, an estimator is efficient if and only if it is asymptotically linear with influence curve $D(P_0)$ equal to the efficient influence curve $D^*(P_0)$:

$$\psi_n - \psi_0 = \frac{1}{n}\sum_{i=1}^{n} D^*(P_0)(O_i) + o_P(1/\sqrt{n}).$$

The efficient influence curve is also called the canonical gradient and is indeed defined as the canonical gradient of the pathwise derivative of the target parameter $\Psi : \mathcal{M} \rightarrow \mathbb{R}$. Specifically, one defines a rich family of one dimensional submodels $\{P(\epsilon) : \epsilon\}$ through P at $\epsilon = 0$, and one represents the pathwise derivative $\frac{d}{d\epsilon}\Psi(P(\epsilon))\big|_{\epsilon=0}$ as an inner product (the covariance operator in the Hilbert space of functions of O with mean zero and inner product $\langle h_1, h_2\rangle = E_P h_1(O)h_2(O))$ $E_P D(P)(O)S(P)(O)$, where $S(P)$ is the score of the path $\{P(\epsilon) : \epsilon\}$, and $D(P)$ is a so called gradient. The unique gradient that is also in the closure of the linear span of all scores generated by the family of one dimensional submodels through P, also called the tangent space at P, is now the canonical gradient $D^*(P)$ at P. Indeed, the canonical gradient can be computed as the projection of any given gradient $D(P)$ onto the tangent space in the Hilbert space $L_0^2(P)$. An interesting result in efficiency theory is that an influence curve of a regular asymptotically linear estimator is a gradient.

In our running example, it can be shown that the efficient influence curve of the additive treatment effect $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ is given by

$$D^*(P_0)(O) = \frac{2A-1}{G_0(A \mid W)}(Y - \bar{Q}_0(A,W)) + \bar{Q}_0(1,W) - \bar{Q}_0(0,W) - \Psi(Q_0). \quad (1)$$

As noted earlier, the influence curve $IC(P_0)$ of an estimator ψ_n also characterizes the limit variance $\sigma_0^2 = P_0 IC(P_0)^2$ of the mean zero normal limit distribution of $\sqrt{n}(\psi_n - \psi_0)$. This variance σ_0^2 can be estimated with $1/n \sum_{i=1}^n IC_n(O_i)^2$, where IC_n is an estimator of the influence curve $IC(P_0)$. Efficiency theory teaches us that for any regular asymptotically linear estimator ψ_n its influence curve has a variance that is larger or equal than the variance of the efficient influence curve, $\sigma_0^{2*} = P_0 D^*(P_0)^2$, which is also called the generalized Cramer-Rao lower bound. In our running example, the asymptotic variance of an efficient estimator is thus estimated with the sample variance of an estimate $D_n^*(O_i)$ of $D^*(P_0)(O_i)$ obtained by plugging in the estimator G_n of G_0 and the estimator $\bar{Q}_n$ of $\bar{Q}_0$, and $\Psi(Q_0)$ is replaced by $\Psi(Q_n) = \frac{1}{n}\sum_{i=1}^n (\bar{Q}_n(1,W_i) - \bar{Q}_n(0,W_i))$.

2.10 Targeted Estimator solves the efficient influence curve equation

The efficient influence curve is a function of O that depends on P_0 through Q_0 and possible a nuisance parameter G_0, and it can be calculated as the canonical gradient of the pathwise derivative of the target parameter mapping along paths through P_0. It is also called the efficient score. Thus, given the statistical model and target parameter mapping, one can calculate the efficient influence curve whose variance defines the best possible asymptotic variance of an estimator, also referred to as the generalized Cramer-Rao lower bound for the asymptotic variance of a regular estimator. The principal building block for achieving asymptotic efficiency of a substitution estimator $\Psi(Q_n)$, beyond Q_n being an excellent estimator of Q_0 as achieved with super-learning, is that the estimator Q_n solves the so called efficient influence curve equation $\sum_{i=1}^n D^*(Q_n, G_n)(O_i) = 0$, for a good estimator G_n of G_0. This property cannot be expected to hold for a super-learner, and that is why the TMLE discussed in Section 4 involves an additional update of the super-learner that guarantees that it solves this efficient influence curve equation.

For example, maximum likelihood estimators solve all score equations, including this efficient score equation that targets the target parameter, but maximum likelihood estimators for large semi parametric models $\mathcal{M}$ typically do not exist for finite sample sizes. Fortunately, for efficient estimation of the target parameter one should only be concerned with solving this particular efficient score tailored for the target parameter. Using the notation $Pf \equiv \int f(o)dP(o)$ for the expectation operator, one way to understand why the efficient influence curve equation indeed targets the true target parameter value is that there are many cases in which $P_0 D^*(P) = \Psi(P_0) - \Psi(P)$, and, in general, as a consequence of $D^*(P)$ being a canonical gradient,

$$P_0 D^*(P) = \Psi(P_0) - \Psi(P) + R(P, P_0), \tag{2}$$

where $R(P, P_0) = o(\parallel P - P_0 \parallel)$ is a term involving second order differences $(P - P_0)^2$. This key property explains why solving $P_0 D^*(P) = 0$ targets $\Psi(P)$ to be close to $\Psi(P_0)$, and thus explains why solving $P_n D^*(Q_n, G_n) = 0$ targets Q_n to fit $\Psi(Q_0)$.

In our running example, we have $R(P, P_0) = R_1(P, P_0) - R_0(P, P_0)$, where $R_a(P, P_0) = \int_w \frac{(G-G_0)(a|W)}{G(a|W)}(\bar{Q} - \bar{Q}_0)(a, W)dP_{W,0}(w)$. So in our example, the remainder $R(P, P_0)$ only involves a cross-product difference $(G - G_0)(\bar{Q} - \bar{Q}_0)$. In particular, the remainder equals zero if either $G = G_0$ or $Q = Q_0$, which is often referred to as double robustness of the efficient influence curve w.r.t. (Q, G) in the causal and censored data literature (see e.g. (81)). This property translates into double robustness of estimators that solve the efficient influence curve estimating equation.

Due to this identity (2), an estimator $\hat{P}$ that solves $P_n D^*(\hat{P}) = 0$, and is in a local neighborhood of P_0 so that $R(\hat{P}, P_0) = o_P(1/\sqrt{n})$, approximately solves $\Psi(\hat{P}) - \Psi(P_0) \approx (P_n - P_0)D^*(\hat{P})$ where the latter behaves a a mean zero centered empirical mean with minimal variance that will be approximately normally distributed. This is formalized in an actual proof of asymptotic efficiency in the next subsection.

2.11 Targeted estimator is asymptotically linear and efficient

In fact, combining $P_n D^*(Q_n, G_n) = 0$ with (2) at $P = (Q_n, G_n)$ yields

$$\Psi(Q_n) - \Psi(Q_0) = (P_n - P_0)D^*(Q_n, G_n) + R_n,$$

where R_n is a second order term. Thus, if second order differences such as $(Q_n - Q_0)^2$, $(Q_n - Q_0)(G_n - G_0)$ and $(G_n - G_0)^2$ converge to zero at a rate faster than $1/\sqrt{n}$, then it follows that $R_n = o_P(1/\sqrt{n})$. To make this assumption as reasonable as possible one should use super-learning for both Q_n and G_n. In addition, empirical process theory teaches us that $(P_n - P_0)D^*(Q_n, g_n) = (P_n - P_0)D^*(Q_0, g_0) + o_P(1/\sqrt{n})$ if $P_0\{D^*(Q_n, G_n) - D^*(Q_0, G_0)\}^2$ converges to zero in probability as n converges to infinity (a consistency condition), and if $D^*(Q_n, G_n)$ falls in a so called Donsker class of functions $O \rightarrow f(O)$ (83). An important Donsker class is the class of all d-variate real valued functions that have a uniform sectional variation norm that is bounded by some universal $M < \infty$: that is, the variation norm of the function itself and the variation norm of its sections are all bounded by this $M < \infty$. This Donsker class condition essentially excludes estimators Q_n, G_n that heavily over fit the data so that their variation norms converge to infinity as n converges to infinity. So under this Donsker class condition, $R_n = o_P(1/\sqrt{n})$, and the consistency condition, we have

$$\psi_n - \psi_0 = \frac{1}{n}\sum_{i=1}^{n} D^*(Q_0, G_0)(O_i) + o_P(1/\sqrt{n}).$$

That is, ψ_n is asymptotically efficient. In addition, the right-hand side converges to a normal distribution with mean zero and variance equal to the variance of the efficient influence curve. So, in spite of the fact that the efficient influence curve equation only represents a finite dimensional equation for an infinite dimensional object Q_n, it implies consistency of $\Psi(Q_n)$ up till a second order term R_n, and even asymptotic efficiency if $R_n = o_P(1/\sqrt{n})$ under some weak regularity conditions.

3 Road map for targeted learning of causal quantity or other underlying full-data target parameters.

This is a good moment to review the roadmap for targeted learning. We have formulated a roadmap for Targeted Learning of a causal quantity that provides a transparent roadmap (47; 35; 37), involving the following steps:

- Defining a full-data model such as a causal model and a parameterization of the observed data distribution in terms of the full-data distribution (e.g., the Neyman-Rubin-Robins counterfactual model (34; 53; 54; 24; 41; 42)) or the structural causal model (35));

- Defining the target quantity of interest as a target parameter of the full-data distribution;

- Establishing identifiability of the target quantity from the observed data distribution under possible additional assumptions that are not necessarily believed to be reasonable;

- Committing to the resulting estimand and the statistical model that is believed to contain the true P_0;

- A sub-roadmap for the TMLE discussed below to construct an asymptotically efficient substitution estimator of the statistical target parameter;

- Establishing an asymptotic distribution and corresponding estimator of this limit distribution to construct a confidence interval;

- Honest interpretation of the results, possibly including a sensitivity analysis (52; 40; 57; 17).

That is, the statistical target parameters of interest are often constructed through the following process. One assumes an underlying model of probability distributions, we will call the full-data model, and one defines the data distribution in terms of this full-data distribution. This can be thought of as modeling: i.e. one obtains a parameterization $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$ for the statistical model $\mathcal{M}$ for some underlying parameter space Θ and parameterization $\theta \rightarrow P_\theta$. The target quantity of interest is defined as some parameter of the full-data distribution, i.e., of θ_0. Under certain assumptions one establishes that the target quantity can be represented as a parameter of the data distribution, a so called estimand: such a result is called an identifiability result for the target quantity. One might now decide to use this estimand as the target parameter and develop a TMLE for this target parameter. Under the non-testable assumptions the identifiability result relied upon, the estimand can be be interpreted as the target quantity of interest, but importantly it can always be interpreted as a statistical feature of the data distribution (due to the statistical model being true), possibly of independent interest. In this manner, one can define estimands that are equal to a causal quantity of interest defined in an underlying (counterfactual) world. The TMLE of this estimand, which is only defined by the statistical model and the target parameter mapping, and thus ignorant of the non-testable assumptions that allowed the causal interpretation of the estimand, provides now an estimator of this causal quantity. In this manner, Targeted Learning is in complete harmony with the development of models such as causal and censored data models and identification results for underlying quantities: the latter just provides us with a definition of a target parameter mapping and statistical model, and thereby the pure statistical estimation problem that needs to be addressed.

4 Targeted Minimum Loss Based Estimation (TMLE)

The TMLE (82; 69; 47) is defined according to the following steps. Firstly, one writes the target parameter mapping as a mapping applied to a part of the data distribution P_0, say $Q_0 = Q(P_0)$, that can be represented as the minimizer of a criterion at the true data distribution P_0 over all candidate values $\{Q(P) : P \in \mathcal{M}\}$ for this part of the data distribution: we refer to this criterion as the risk $R_{P_0}(Q)$ of the candidate value Q.

Typically, the risk at a candidate parameter value Q can be defined as the expectation under the data distribution of a loss function $(O, Q) \mapsto L(Q)(O)$ that maps the unit data structure and the candidate parameter value in a real value number: $R_{P_0}(Q) = E_{P_0} L(Q)(O)$. Examples of loss functions are the squared error loss for a conditional mean and the log-likelihood loss for a (conditional) density. This representation of Q_0 as a minimizer of a risk allows us to estimate it with (e.g., loss-based) super-learning.

Secondly, one computes the efficient influence curve $(O, P) \mapsto D^*(Q(P), G(P))(O)$ identified by the canonical gradient of the pathwise derivative of the target parameter mapping along paths through a data distribution P, where this efficient influence curve does only depend on P through $Q(P)$ and some nuisance parameter $G(P)$. Given an estimator G_n, one now defines a path $\{Q_{n,G_n}(\epsilon) : \epsilon\}$

with Euclidean parameter ϵ through the super-learner Q_n whose score

$$\frac{d}{d\epsilon} L(Q_{n,G_n}(\epsilon))\Big|_{\epsilon=0}$$

at $\epsilon = 0$ spans the efficient influence curve $D^*(Q_n, G_n)$ at the initial estimator (Q_n, G_n): this is called a least-favorable parametric submodel through the super-learner.

In our running example, we have $Q = (\bar{Q}, Q_W)$ so that it suffices to construct a path through $\bar{Q}$ and Q_W with corresponding loss functions and show that their scores span the efficient influence curve (1. We can define the path $\{\bar{Q}_G(\epsilon) = \bar{Q} + \epsilon C(G) : \epsilon\}$, where $C(G)(O) = (2A-1)/G(A \mid W)$ and loss function $L(\bar{Q})(O) = -\{Y \log \bar{Q}(A,W) + (1-Y)\log(1-\bar{Q}(A,W))\}$. Note that

$$\frac{d}{d\epsilon} L(\bar{Q}_G(\epsilon))(O)\Big|_{\epsilon=0} = D^*_Y(\bar{Q}, G) = \frac{2A-1}{G(A \mid W)}(Y - \bar{Q}(A,W)).$$

We also define the path $Q_W(\epsilon) = (1 + \epsilon D^*_W(\bar{Q}, Q_W))Q_W$ with loss-function $L(Q_W)(W) = -\log Q_W(W)$, where $D^*_W(Q)(O) = \bar{Q}(1,W) - \bar{Q}(0,W) - \Psi(Q)$. Note that

$$\frac{d}{d\epsilon} L(Q_W(\epsilon))\Big|_{\epsilon=0} = D^*_W(Q).$$

Thus, if we define the sum loss function $L(Q) = L(\bar{Q}) + L(Q_W)$, then

$$\frac{d}{d\epsilon} L(\bar{Q}_G(\epsilon), Q_W(\epsilon))\Big|_{\epsilon=0} = D^*(Q, G).$$

This proves that indeed these proposed paths through $\bar{Q}$ and Q_W and corresponding loss-functions span the efficient influence curve $D^*(Q,G) = D^*_W(Q) + D^*_Y(\bar{Q}, G)$ at (Q, G), as required.

The dimension of ϵ can be selected to be equal to the dimension of the target parameter ψ_0, but by creating extra components in ϵ one can arrange to solve additional score equations beyond the efficient score equation, providing important additional flexibility and power to the procedure. In our running example, we can use an ϵ_1 for the path through $\bar{Q}$ and a separate ϵ_2 for the path through Q_W. In this case, the TMLE update Q^*_n will solve two score equations $P_n D^*_W(Q^*_n) = 0$ and $P_n D^*_Y(\bar{Q}^*_n, G_n) = 0$, and thus, in particular, $P_n D^*(Q^*_n, G_n) = 0$. In this example, the main benefit of using a bivariate $\epsilon = (\epsilon_1, \epsilon_2)$ is that the TMLE does not update $Q_{W,n}$ (if selected to be the empirical distribution) and converges in a single step.

One fits the unknown parameter ϵ of this path by minimizing the empirical risk $\epsilon \to P_n L(Q_{n,G_n}(\epsilon))$ along this path through the super-learner, resulting in an estimator ϵ_n. This defines now an update of the super-learner fit defined as $Q^1_n = Q_{n,G_n}(\epsilon_n)$. This updating process is iterated till $\epsilon_n \approx 0$. The final update we will denote with Q^*_n, the TMLE of Q_0, and the target parameter mapping applied to Q^*_n defines the TMLE of the target parameter ψ_0. This TMLE Q^*_n solves the efficient influence curve equation $\sum_{i=1}^n D^*(Q^*_n, G_n)(O_i) = 0$, providing the basis (in combination with statistical properties of (Q^*_n, G_n) for establishing that the TMLE $\Psi(Q^*_n)$ is asymptotically consistent, normally distributed and asymptotically efficient, as shown above.

In our running example, we have $\epsilon_{1n} = \arg\min_\epsilon P_n L(\bar{Q}^0_{n,G_n}(\epsilon))$, while $\epsilon_{2n} = \arg\min P_n L(Q_{W,n}(\epsilon_2))$ equals zero. That is, the TMLE does not update $Q_{W,n}$ since the empirical distribution is already a nonparametric maximum likelihood estimator solving all score equations. In this case $\bar{Q}^*_n = \bar{Q}^1_n$ since the convergence of the TMLE-algorithm occurs in one step, and, of course, $Q^*_{W,n} = Q_{W,n}$ is just the initial empirical distribution function of $W_1, \ldots, W_n$. The TMLE of ψ_0 is the substitution estimator $\Psi(Q^*_n)$.

5 Advances in Targeted Learning

As apparent from the above presentation, TMLE is a general method that can be developed for all types of challenging estimation problems. It is a matter of representing the target parameters as a parameter of a smaller Q_0, defining a path and loss-function with generalized score that spans the efficient influence curve, and the corresponding iterative targeted minimum loss-based estimation algorithm.

We have used this framework to develop TMLE in a large number of estimation problems that assumes that $O_1, \ldots, O_n \sim_{iid} P_0$. Specifically, we developed TMLE of a large variety of effects (e.g., causal) of a single and multiple time point interventions on an outcome of interest that may be subject to right-censoring, interval censoring, case-control sampling, and time-dependent confounding: see e.g. (4; 69; 45; 46; 48; 29; 30; 31; 3; 33; 39; 49; 77; 62; 20; 50; 85; 12; 13; 15; 14; 16; 63; 70; 64; 28; 23; 22; 36; 7; 6; 56; 32; 27; 26; 65; 25; 87; 90; 89; 56; 8).

An original example of a particular type of TMLE (based on a double robust parametric regression model) for estimation of a causal effect of a point-treatment intervention was presented in (58) and we refer to (50) for a detailed review of this earlier literature and its relation to TMLE.

It is beyond the scope of this overview article to get into a review of some of these examples. For a general comprehensive book on Targeted Learning, which includes many of these applications on TMLE and more, we refer to

(47).

To provide the reader with a sense, consider generalizing our running example to a general longitudinal data structure $O = (L(0), A(0), \ldots, L(K), A(K), Y)$, where $L(0)$ are baseline covariates, $L(k)$ are time dependent covariates realized between intervention nodes $A(k-1)$ and $A(k)$, and Y is the final outcome of interest. The intervention nodes could both include censoring variables as well as treatment variables: the desired intervention for the censoring variables is always ”no censoring” since the outcome Y is only of interest when it is not subject to censoring (in which case it might just be a forward imputed value, for example).

One may now assume a structural causal model of the type discussed earlier and be interested in the mean counterfactual outcome under a particular intervention on all the intervention nodes, where these interventions could be static, dynamic or even stochastic. Under the so called sequential randomization assumption, this target quantity is identified by the so called G-computation formula for the post-intervention distribution corresponding with a stochastic intervention g^*:

$$\begin{aligned} P_{0,g^*}(O) &= \prod_{k=0}^{K+1} P_{0,L(k)\mid\bar{L}(k-1),\bar{A}(k-1)}(L(k) \mid \bar{L}(k-1), \bar{A}(k-1)) \\ & \prod_{k=0}^{K} g_k^*(A(k) \mid \bar{A}(k-1), \bar{L}(k)). \end{aligned}$$

Note that this post-intervention distribution is nothing else than the actual distribution of O factorized according to the time-ordering but with the true conditional distributions of $A(k)$, given $\bar{L}(k), \bar{A}(k-1))$, replaced by the desired stochastic intervention. The statistical target parameter is thus $E_{P_{0,g^*}} Y_{g^*}$, i.e., the mean outcome under this post-intervention distribution. A big challenge in the literature has been to develop robust efficient estimators of this estimand, and, more generally, one likes to estimate this mean outcome under a user supplied class of stochastic interventions g^*. Such robust efficient substitution estimators have now been developed using the TMLE framework (23; 36), where the latter is a TMLE inspired by important double robust estimators established in earlier work of (2). This work thus includes causal effects defined by working marginal structural models for static and dynamic treatment regimens, time to event outcomes, incorporating right-censoring,

In many data sets one is interested in assessing the effect of one variable on an outcome, controlling for many other variables, across a large collection of variables. For example, one might want to know the effect of a single nucleotide polymorphism (SNP) on a trait of a subject across a whole genome, controlling each time for a large collection of other SNP's in the neighborhood of the SNP in question. Or one is interested in assessing the effect of a mutation in the HIV virus on viral load drop (measure of drug resistance) when treated with a particular drug class, controlling for the other mutations in the HIV virus and for characteristics of the subject in question. Therefore, it is important to carefully define the effect of interest for each variable. If the variable is binary, one could use the target parameter $\Psi(P) = E_P\{E_P(Y \mid A = 1, W) - E_P(Y \mid A = 0, W)\}$ in our running example, but with A now being the SNP in question and W being the variables one wants to control for, while Y is the outcome of interest. We often refer to such a measure as a particular variable importance measure. Of course, one now defines such a variable importance measure for each variable. When the variable is continuous, the above measure is not appropriate. In that case, one might define the variable importance as the projection of $E_P(Y \mid A, W) - E_P(Y \mid A = 0, W)$ onto a linear model such as βA, and use β as the variable importance measure of interest (9), but one could think of a variety of other interesting effect measures. Either way, for each variable, one uses a TMLE of the corresponding variable importance measure. The stacked TMLE across all variables is now an asymptotically linear estimator of the stacked variable importance measure with stacked influence curve, and thus approximately follows a multivariate normal distribution that can be estimated from the data. One can now carry out multiple testing procedures controlling a desired family wise type I error rate and construct simultaneous confidence intervals for the stacked variable importance measure, based on this multivariate normal limit distribution. In this manner, one uses targeted learning to target a large family of target parameters while still providing honest statistical inference taking into account multiple testing. This approach deals with a challenge in machine learning in which one wants estimators of a prediction function that simultaneously yield good estimates of the variable importance measures. Examples of such efforts are random forest and LASSO, but both regression methods fail to provide reliable variable importance measures, and fail to provide any type of statistical inference. The truth is that if the goal is not prediction but to obtain a good estimate of the variable importance measures across the variables, then one should target the estimator of the prediction function towards the particular variable importance measure, for each variable separately, and only then one obtains valid estimators and statistical inference. For TMLE of effects of variables across a large set of variables, a so called variable importance analysis, including the application to genomic data sets we refer to: (66; 67; 68; 85; 86; 9).

Software has been developed in the form of general R-packages implementing super-learning and TMLE for general longitudinal data structures: these packages are publicly available on CRAN under the function names tmle(), ltmle(), and superlearner().

Beyond the development of TMLE in this large variety of complex statistical estimation problems, as usual, the careful study of real world applications resulted in new challenges for the TMLE, and in response to that we have developed general TMLE that have additional properties dealing with these challenges. In particular, we have shown that TMLE has the flexibility and capability to enhance the finite sample performance of TMLE under the following specific challenges that come with real data applications.

Dealing with rare outcomes: If the outcome is rare, then the data is still sparse even though the sample size might be quite large. When the data is sparse w.r.t. the question of interest, the incorporation of global constraints of the statistical model becomes extremely important and can make a real difference in a data analysis. Consider our running example and suppose that Y is the indicator of a rare event. In such cases it is often known that the probability of $Y = 1$, conditional on a treatment and covariate configuration, should not exceed a certain value $\delta > 0$: e..g, the marginal prevalence is known and it is known that there are no subpopulations that increase the relative risk by more than a certain factor relative marginal prevalence. So the statistical model should now include the global constraint that $\bar{Q}_0(A, W) < \delta$ for some known $\delta > 0$. A TMLE should now be based on an initial estimator $\bar{Q}_n$ satisfying this constraint, and the least favorable submodel $\{\bar{Q}_{n,G_n}(\epsilon) : \epsilon\}$ should also satisfy this constraint for each ϵ so that it is a real *submodel*. In (1) such a TMLE is constructed and it is demonstrated to very significantly enhance its practical performance for finite sample sizes. Even though a TMLE ignoring this constraint would still be asymptotically efficient, by ignoring this important knowledge, its practical performance for finite samples suffers.

Targeted Estimation of nuisance parameter G_0 in TMLE: Even though an asymptotically consistent estimator G_n of G_0 yields an asymptotically efficient TMLE, the practical performance of the TMLE might be enhanced by tuning this estimator G_n not only w.r.t. to its performance in estimating G_0, but also w.r.t. how well the resulting TMLE fits ψ_0. Consider our running example. Suppose that among the components of W there is a W_j that is an almost perfect predictor of A, but has no effect on the outcome Y. Inclusion of such a covariate W_j in the fit of G_n makes sense if the sample size is very large and one tries to remove some residual confounding due to not adjusting for W_j, but in most finite samples adjustment for W_j in G_n will hurt the practical performance of TMLE, and effort should be put in variables that are stronger confounders than W_j. We developed a method for building an estimator G_n that uses as criterion the change in fit between initial estimator of Q_0 and the updated estimator (i.e., the TMLE), and thereby selects variables that result in the maximal increase in fit during the TMLE updating step. However, eventually, as sample size converges to infinity, all variables will be adjusted for so that asymptotically the resulting TMLE is still efficient. This version of TMLE is called the collaborative TMLE since it fits G_0 in collaboration with the initial estimator Q_n (47; 77; 23; 62; 20). Finite sample simulations and data analyses have shown remarkable important finite sample gains of C-TMLE relative to TMLE (see above references).

Cross-validated TMLE: The asymptotic efficiency of TMLE relies on a so called Donsker class condition. For example, in our running example, it requires that $\bar{Q}_n$ and G_n are not too erratic functions of (A, W). This condition is not just theoretical but one can observe its effects in finite samples by evaluating the TMLE when using a heavily over fitted initial estimator. This makes sense, since if we use an over fitted initial estimator, there is little reason to think that the ϵ_n that maximizes the fit of the update of the initial estimator along the least favorable parametric model will still do a good job. Instead, one should use the fit of ϵ that maximizes a honest estimate of the fit of the resulting update of the initial estimator as measured by the cross-validated empirical mean of the loss function. This insight results in a so called cross-validated TMLE, and we have proven that one can establish asymptotic linearity of this CV-TMLE *without* a Donsker class condition (88; 47);: thus the CV-TMLE is asymptotically linear under weak conditions than the TMLE.

Guaranteed minimal performance of TMLE: If the initial estimator Q_n is inconsistent, but G_n is consistent, then the TMLE is still consistent for models and target parameters in which the efficient influence curve is double robust. However, there might be other estimators that will now asymptotically beat the TMLE, since the TMLE is not efficient anymore. The desire for estimators to have a guarantee to beat certain user supplied estimators was formulated and implemented for double robust estimating equation based estimators in (51). Such a property can also be arranged within the TMLE framework by incorporating additional fluctuation parameters in its least favorable submodel though the initial estimator so that the TMLE solves additional score equations that guarantee that it beats a user supplied estimator, even under heavy misspecification of the initial estimator Q_n (23; 25).

Targeted selection of initial estimator in TMLE: In situations where it is unreasonable to expect that the initial estimator Q_n will be close to the true Q_0, such as in randomized controlled trials in which the sample size is small, one may improve the efficiency of the TMLE by using a criterion for tuning the initial estimator that directly evaluates the efficiency of the resulting TMLE of ψ_0. This general insight was formulated as empirical efficiency maximization in (55) and further worked out in the TMLE context in chapter 12 and Appendix of (47).

Double robust inference: If the efficient influence curve is double robust, then the TMLE remains consistent if either Q_n or G_n is consistent. However, if one uses a data adaptive consistent estimator of G_0 (and thus with bias larger than $1/\sqrt{n}$)), and Q_n is inconsistent, then the bias of G_n might directly map into a bias for the resulting TMLE of ψ_0 of the same order. As a consequence, the TMLE might have a bias w.r.t. ψ_0 that is larger than $O(1/\sqrt{n})$, so that it is not asymptotically linear. However, one can incorporate additional fluctuation parameters in the least-favorable sub-model (by also fluctuating G_n) to guarantee that the TMLE remains asymptotically linear with known influence curve when either Q_n or G_n is inconsistent, but we do not know which one (72). So these enhancements of TMLE result in TMLE that are asymptotically linear under weaker conditions than a standard TMLE, just like the CV-TMLE removed a condition for asymptotic linearity. These TMLE now involve not only targeting Q_n but also targeting G_n to guarantee that when Q_n is misspecified the required smooth function of G_n will behave as a TMLE, and if G_n is misspecified, that the required smooth functional of Q_n is still asymptotically linear. The same method was used to develop an IPTW-estimator that targets G_n so that the IPTW-estimator is asymptotically linear with known influence curve even when the initial estimator of G_0 is estimated with a highly data adaptive estimator.

Super-learning based on CV-TMLE of the conditional risk of a candidate estimator: Super-learner relies on a cross-validated estimate of the risk of a candidate estimator. The oracle inequalities of the cross-validation selector assumed that the cross-validated risk is simply an empirical mean over the validation sample of a loss function at the candidate estimator based on training sample, averaged across different sample splits, where we generalized these results to loss functions that depend on an unknown nuisance parameter (which are thus estimated in the cross-validated risk).

For example, suppose that in our running example A is continuous, and we are concerned with estimation of the dose-response curve $(E_0 Y_a : a)$, where

$E_0 Y_a = E_0 E_0(Y \mid A = a, W)$. One might define the risk of a candidate dose response curve as a mean squared error w.r.t. the true curve $E_0 Y_a$. However, this risk of a candidate curve is itself an unknown real valued target parameter. Contrary, to standard prediction or density estimation, this risk is not simply a mean of a known loss function, and, the proposed unknown loss functions indexed by a nuisance parameter can have large values making the cross-validated risk a non robust estimator. Therefore, we have proposed to estimate this conditional risk of candidate curve with TMLE and, similarly, the conditional risk of a candidate estimator with a CV-TMLE. One can now develop a super-learner that uses CV-TMLE as an estimate of the conditional risk of a candidate estimator (79; 16). We applied this to construct a super-learner of the causal dose response curve for a continuous valued treatment, and we obtained a corresponding oracle inequality for the performance of the cross-validation selector (16).

6 Eye on the future of Targeted Learning

We hope that the above clarifies that targeted learning is an ongoing exciting research area that is able to address important practical challenges. Each new application concerning learning from data can be formulated in terms of a statistical estimation problem with a large statistical model and a target parameter. One can now use the general framework of super-learning and TMLE to develop efficient targeted substitution estimators and corresponding statistical inference. As is apparent from the previous section, the general structure of TMLE and super-learning appears to be flexible enough to handle/adapt to any new challenges that come up, allowing researchers in targeted learning to make important progress in tackling real world problems. By being honest in the formulation, typically new challenges come up asking for expert input from a variety of researchers, ranging from subject-matter scientists, computer scientists, to statisticians. Targeted Learning requires multi-disciplinary teams, since it asks for careful knowledge about data experiment, the questions of interest, possible informed guesses for estimation that can be incorporated as candidates in the library of the super-learner, and input from the state of the art in computer science to produce scalable software algorithms implementing the statistical procedures.

There are a variety of important areas of research in Targeted Learning we began to explore.

Variance estimation: The asymptotic variance of an estimator such as the TMLE, i.e. the variance of the influence curve of the estimator, is just another target parameter of great interest. It is common practice to estimate this asymptotic variance with an empirical sample variance of the estimated influence curves. However, in the context of sparsity, influence curves can be large, making such an estimator highly non robust. In particular, such a sample mean type estimator will not respect the global constraints of the model. Again, this is not just a theoretical issue since we have observed that in sparse data situations standard estimators of the asymptotic variance often under estimate the variance of the estimator, thereby, resulting in overly optimistic confidence intervals. This sparsity can be due to rare outcomes or strong confounding or highly informative censoring, for example, and naturally occurs even when sample sizes are large. Careful inspection of these variance estimators shows that the essential problem is that these variance estimators are not substitution estimators. Therefore, we are in the process to apply TMLE to improve the estimators of the asymptotic variance of TMLE of a target parameter, thereby improving the finite sample coverage of our confidence intervals, especially in sparse-data situations.

Dependent data: Contrary to experiments that involve random sampling from a target population, if one observes the real world over time, then naturally there is no way to argue that the experiment can be represented as a collection of independent experiments, let alone, identical independent experiments. An environment over time and space is a single organism that cannot be separated out into independent units without making very artificial assumptions and losing very essential information: the world needs to be seen as a whole to see truth. Data collection in our societies is moving more an more towards measuring total populations over time, resulting in what we often refer to as Big Data, and these populations consists of interconnected units. Even in randomized controlled settings where one randomly samples units from a target population, one often likes to look at the past data and change the sampling design in response to the observed past, in order to optimize the data collection w.r.t. certain goals. Once again, this results in a sequence of experiments that cannot be viewed as independent experiments, the next experiment is only defined once one knows the data generated by the past experiments.

Therefore we believe that our models that assume independence, even though they are so much larger than the models used in current practice, are still not realistic models in many applications of interest. On the other hand, even when the sample size equals 1, things are not hopeless if one is willing to assume that the likelihood of the data factorizes in many factors due to conditional independence assumptions, and stationarity assumptions that state that conditional distributions might be constant across time or that different units are subject tot the same laws for generating their data as a function of their parent variables. In more recent research we have started to develop TMLE for statistical models that do not assume that the unit-specific data structures are independent, handling adaptive pair matching in community randomized controlled trials, group sequential adaptive randomization designs, and studies that collect data on units that are interconnected through a causal network (69; 11; 10; 74; 71).

Data adaptive target parameters: It is common practice that people first look at data before determining their choice of target parameter they want to learn, even though it is taught that this is unacceptable practice since it makes the p-values and confidence intervals unreliable. But, may be we should view this common practice as a sign that a priori specification of the target parameter (and null hypothesis) limits the learning from data too much, and by enforcing it we only force data analysts to cheat. Current teaching would tell us that one is only allowed to do this by splitting the sample, use one part of the sample to generate a target parameter, and use the other part of the sample to estimate this target parameter and obtain confidence intervals. Clearly, this means that one has to sacrifice a lot of sample size for being allowed to look at the data first. Another possible approach for allowing to obtain inference for a data driven parameter is to a priori formulate a large class of target parameters, and use multiple testing or simultaneous confidence interval adjustments. However, also with this approach one has to pay a big price through the multiple testing adjustment and one still needs to a priori list the target parameters.

For that purpose, acknowledging that one likes to mine the data to find interesting questions that are supported by the data, we developed statistical inference based on CV-TMLE for a large class of target parameters that are defined as functions of the data (78). This allows one to define an algorithm that when applied to the data generates an interesting target parameter, while we provide formal statistical inference in terms of confidence intervals for this data adaptive target parameter. This provides a much broader class of a priori specified statistical analyses than current practice which requires a priori specification of the target parameter, while still providing valid statistical inference. We believe that this is a very promising direction for future research, opening up many new applications which would normally be overlooked.

Optimal individualized treatment: One is often interested in learning the best rule for treating a subject in response to certain time-dependent measurements on that subject, where best rule might be defined as the rule that optimizes the expected outcome. Such a rule is called an individualized treatment rule, or dynamic treatment regimen, and an optimal treatment rule is defined as the rule that minimizes the mean outcome for a certain outcome (e.g. indicator of death or other health measurement). We started to address data adaptive learning of the best possible treatment rule by developing super-learners of this important target parameter, while still providing statistical inference (and thus confidence intervals) for the mean of the outcome in the counterfactual world in which one applies this optimal dynamic treatment to everybody in the target population (73). In particular, this problem itself provides a motivation for a data adaptive target parameter, namely the mean outcome under a treatment rule fitted based on the data. Optimal dynamic treatments has been an important area in statistics and computer science, but we target this problem within the framework of targeted learning, thereby avoiding reliance on unrealistic assumptions that cannot be defended and will heavily affect the true optimality of the fitted rules.

Statistical inference based on higher order inference: Another key assumption the asymptotic efficiency or asymptotic linearity of TMLE relies upon is that the remainder/second order term $R_n = o_P(1/\sqrt{n})$. For example, in our running example this means that the product of the rate at which the super-learner estimators of Q_0 and G_0 converge to their target converges to zero at a faster rate than $1/\sqrt{n}$. The density estimation literature proves that if the density is many times differentiable, then it is possible to construct density estimators whose bias is driven by the last term of a higher order Tailor expansion of the density around a point. Robins, van der Vaart, and colleagues (43) have developed theory based on higher order influence functions, under the assumption that the target parameter is higher order pathwise differentiable. Just as density estimators exploiting underlying smoothness, this theory also aims to construct estimators of higher order pathwise differentiable target parameters whose bias is driven by the last term of the higher order Tailor expansion of the target parameter. The practical implementation of the proposed estimators have been challenging and is suffering from a lack of robustness. Targeted learning based on these higher order expansions (thus not only incorporating the first order efficient influence function but also the higher order influence functions that define the Tailor expansion of the target parameter) appears to be a natural area of future research to further build on these advances.

Online TMLE: Trading off statistical optimality and computing cost.

We will be more and more confronted with online data bases that continuously grow and are massive in size. Nonetheless, one wants to know if the new data changes the inference about target parameters of interest, and one wants to know it right away. Recomputing the TMLE based on the old data augmented with the new chunk of data would be immensely computer intensive. Therefore, we are confronted with the challenge on constructing an estimator that is able to update a current estimator without having to recompute the estimator, but instead one wants to update it based on computations with the new data only. More generally, one is interested in high quality statistical procedures that are scalable. We started doing research in such online TMLE that preserve all or most of the good properties of TMLE but can be continuously updated where the number of computations required for this update is only a function of the size of the new chunk of data.

7 Historical Philosophical Perspective on Targeted Learning: A reconciliation with machine learning

In the previous sections the main characteristics of TMLE/SL methodology have been outlined. We introduced the most important fundamental ideas and statistical concepts, urged the need for revision of current data-analytic practice and showed some recent advances and application areas. Also research in progress on such issues as dependent data and data adaptive target parameters has been brought forward. In this section we put the methodology in a broader historical-philosophical perspective, trying to support the claim that its relevance exceeds the realms of statistics in a strict sense, and even those of methodology. To this aim we will discuss both the significance of TMLE/SL for contemporary epistemology and its implications for the current debate on Big Data and the generally advocated, emerging new discipline of Data Science. Some of these issues have been elaborated more extensively in (59; 44; 60; 61) where we have put the present state of statistical data analysis in a historical and philosophical perspective with the purpose to clarify, understand and account for the current situation in statistical data analysis and relate the main ideas underlying TMLE/SL to it.

First and foremost, it must be emphasized that rather than extending the toolkit of the data-analyst, TMLE/SL establishes a new methodology. From a technical point of view it offers an integrative approach to data-analysis or statistical learning by combining inferential statistics with techniques derived from the field of computational intelligence. This field includes such related and usually eloquently phrased disciplines like machine learning, data mining, knowledge discovery in databases and algorithmic data analysis. 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. According to TMLE/SL all these elements should be related to or defined in terms of (properties of) the data generating distribution and to this aim the methodology provides both clear heuristics and formal underpinnings. Among other things this means that the concept of a statistical model is reestablished in a prudent and 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 leading role in any theory of / approach to learning from data, whether it deals with establishing causal relations, classifying or clustering, time series forecasting or multiple testing. Indeed, inferential statistics arose at the background of randomness and variation in a world represented or encoded by probability distributions, and it has therefore always presumed and exploited the sample-population dualism, that underlies the very idea of estimation. Nevertheless, the whole concept of estimation seems to be discredited and disregarded in contemporary data analytical practice.

In fact, 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 of these can be traced back to Karl Pearsons goodness-of-fit-approach to data-analysis or to the Fisherian tradition of significance testing and ML- estimation. Some principles and techniques have been derived from the Neyman-Pearson school of hypothesis testing, such as the comparison between two alternative hypothesis and the identification of two kinds of errors of usual unequal importance that should be dealt with. And, last but not least, the toolkit contains all kinds of ideas taken from the Bayesian paradigm, that rigorously pulls statistics into the realms of epistemology. We only have to refer here to the subjective interpretation of probability and the idea that hypotheses should be analyzed in a probabilistic way by assigning probabilities to these hypotheses, thus abandoning the idea that the parameter is a fixed, unknown quantity and thus moving the knowledge about the hypotheses from the meta-language into the object language of probability calculus. In spite of all this, the burgeoning statistical textbook market offers many primers and even advanced studies, that wrongly suggest a uniform and united field with foundations that are fixed, and on which full agreement has been reached. It offers a toolkit based on the alleged unification of ideas and methods derived from the aforementioned traditions. As pointed out in (59) this situation is rather inconvenient from a philosophical point of view 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, this transition became recognizable in old, established sciences like physics (kinetic gas theory, statistical mechanics of Bolzmann, Maxwell and Gibbs), but especially in new emerging disciplines like the social sciences (Quetelet and later Durkheim), 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 many 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 on the relation between knowledge and reality like: What is reality? Does it exist mind-independent? Do we have access to it? If yes, how? Do our postulated theoretical entities exist? How do they correspond to reality? Can we make true statements about it? If yes, what is truth and how is it connected to reality? The analyses conducted to address 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 natural laws. All too often scientific reasoning seems nearly synonymous with probabilistic reasoning. In view of the fact that scientific inference more and more depends on probabilistic reasoning and that statistical analysis is not as well-founded as might be expected, the issue addressed in this chapter is of crucial importance for epistemology (59).

Despite these philosophical objections against the hybrid character of inferential statistics, its successes were enormous in the first decades of the twentieth century. In newly established disciplines like psychology and economics significance testing and maximum likelihood estimation were applied with methodological rigor in order to enhance prestige and apply scientific method to their field. Although criticism that a mere chasing of low p-values and naive use of parametric statistics did not do justice to specific characteristics of the sciences involved, emerged from the start of the application of statistics, the success story was immense. However, this rise of the inference experts, like Gigerenzer calls them in The Rise of Statistical Thinking, was just a phase or stage in the development of statistics and data analysis, which manifests itself as a Hegelian triptych, that unmistakably is now being completed in the era of Big Data. After this thesis of a successful, but ununified field of inferential statistics, an antithesis in the Hegelian sense of the word was unavoidable and it was this antithesis that gave rise to the current situation in data-analytical practice as well. Apart from the already mentioned Bayesian revolt, the rise of non-parametric statistics in the thirties must be mentioned here as an intrinsically statistical criticism that heralds this antithesis. The major caesura in this process however was the work of John Tukey in the sixties and seventies of the previous century. After a long career in statistics and other mathematical disciplines Tukey wrote Explorative Data analysis in 1978. This study is in many ways a remarkable, unorthodox book. First, it contains no axioms, theorems , lemmas or proofs, and even barely formulas. There are no theoretical distributions, significance tests, p-values, hypothesis tests, parameter estimation and confidence intervals. No inferential or confirmatory statistics, but just the understanding of data, looking for patterns, relationships and structures in data, and visualizing the results. According to Tukey the statistician is a detective; as a contemporary Sherlock Holmes he must strive for signs and " clues ". Tukey maintains this metaphor consistently throughout the book and wants provide the data analyst with a toolbox full of methods for understanding frequency distributions, smoothing techniques, scale transformations , and above all, many graphical techniques for exploration, storage, and summary illustrations of data. The unorthodox approach of Tukey in EDA reveals not so much a contrarian spirit, but rather a fundamental dissatisfaction with the prevailing statistical practice and the underlying paradigm of inferential / confirmatory statistics (60).

In EDA Tukey endeavors to emphasize the importance of confirmatory, classical statistics, but this looks for the main part a matter of politeness and courtesy. In fact, he had already put his cards on the table in 1962 in the famous opening passage from The future of data analysis: For a long time I have thought that I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt. And when I have pondered about why such techniques as the spectrum analysis of time series have proved so useful, it has become clear that their 'dealing with fluctuations' aspects are, in many circumstances, of lesser importance than the aspects that would already have been required to deal effectively with the simpler case of very extensive data where fluctuations would no longer be a problem. All in all, I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of mathematical statistics which apply to analyzing data. (...) Data analysis is a larger and more varied field than inference, or allocation. Also in other writings Tukey makes a sharp distinction between statistics and data analysis.

First, Tukey gave unmistakable an impulse to the emancipation of the descriptive / visual approach, after pioneering work of William Playfair (18th century) and Florence Nightingale (19th century) on graphical techniques, that were soon overshadowed by the " inferential " coup, which marked the probabilistic revolution. Furthermore, it is somewhat ironic that many consider Tukey a pioneer of computational fields such as data mining and machine learning, although he himself preferred a small role for the computer in his analysis and kept it in the background . More importantly, however, because of his alleged anti - theoretical stance, Tukey is sometimes considered the man who tried to reverse or undo the Fisherian revolution, and an exponent or forerunner of today's " erosion of models, the view that all models are wrong, the classical notion of truth is obsolete and pragmatic criteria as predictive success in data analysis must prevail. Also the idea, currently frequently uttered in the data analytical tradition that the presence of big data will make much of the statistical machinery superfluous is an import aspect of the here very briefly sketched antithesis. Before we come to the intended synthesis, the final stage of the Hegelian triptych, let us make two remarks concerning Tukeys heritage. Although it almost sounds like a cliché, yet it must be noted that EDA techniques nowadays are routinely applied in all statistical packages along with in itself sometimes hybrid inferential methods. In the current empirical methodology EDA is integrated with inferential statistics at different stages of the research process. Secondly, it could be argued that Tukey did not so much undermine the revolution initiated by Galton and Pearson, but understands the ultimate consequences of it. It was Galton who had shown that variation and change are intrinsic in nature, that we have to look was looking for the deviant, the special or the peculiar. Is was Pearson who did realize that the constraints of the normal distribution (Laplace , Quetelet) had to be abandoned and who distinguished different families of distributions as an alternative. Galton 's heritage was just slightly under pressure hit by the successes of the parametric Fisherian statistics on strong model assumptions and it could well be stated that this was partially reinstated by Tukey .

Unsurprisingly, the final stage of the Hegelian triptych strives for some convergence if not synthesis. The 19th century dialectical German philosopher G.F.W. Hegel argued that history is a process of becoming or development, in which a thesis evokes and binds itself to an antithesis; in addition both are placed at a higher level to be completed and to result in a fulfilling synthesis. Applied to the less metaphysically oriented present problem, this dialectical principle seems particularly relevant in the era of Big Data, which makes a reconciliation between inferential statistics and computational science imperative. Big data sets high demands and offers challenges to both. For example, it sets high standards for data management, storage and retrieval and has great influence on the research of efficiency of machine learning algorithms. But it is also accompanied by new problems, pitfalls and challenges for statistical inference and its underlying mathematical theory . Examples include the effects of wrongly -specified models, the problems of small, high-dimensional datasets (microarray data), the search for causal relationships in non - experimental data, quantifying uncertainty, efficiency theory, et cetera. The fact that many data-intensive empirical sciences are highly dependent on machine learning algorithms and statistics makes bridging the gap of course, for practical reasons compelling.

In addition, it seems that Big Data itself also transforms the nature of knowledge: the way of acquiring knowledge, research methodology, nature and status of models and theories. In the reflections of all the briefly sketched contradiction often emerges and in the popular literature the differences are usually enhanced, leading to annexation of Big Data by one of the two disciplines.

Of course the gap between both has many aspects, both philosophical and technical that have been left out here. However, it must be emphasized that for the main part Targeted Learning intends to support the reconciliation between inferential statistics and computational intelligence. It starts with the specification of a non - parametric and semi- parametric model that contains only the realistic background knowledge and focuses on the parameter of interest, which is considered as a property of the as yet unknown, true data -generating distribution. From a methodological point of view it is a clear imperative that model and parameter of interest must be specified in advance. The (empirical) research question must be translated in terms of the parameter of interest and a rehabilitation of the concept model is achieved. Then, Targeted Learning involves a flexible, data-adaptive estimation procedure that proceeds in two steps. First an initial estimate is searched on the basis of the relevant part of the true distribution that is needed to evaluate the target parameter. This initial estimator is found by means of the super learning- algorithm. In short, this is based on a library of many diverse analytical techniques ranging from logistic regression to ensemble techniques, random forest and support vector machines. Because the choice of one of these techniques by human intervention is highly subjective and the variation in the results of the various techniques usually substantial, SL uses a sort of weighted sum of the values calculated by means of cross-validation. Based on these initial estimator, the second stage of the estimation procedure can be initiated. The initial fit is updated with the goal of an optimal bias -variance trade-off for the parameter of interest. This is accomplished with a targeted maximum likelihood estimator of the fluctuation parameter of a parametric sub-model selected by the initial estimator. The statistical inference is then completed by calculating standard errors on the basis of " influence - curve theory" or resampling techniques. This parameter estimation retains a crucial place in the data analysis . If one wants to do justice to variation and change in the phenomena, then you cannot deny Fishers unshakable insight that randomness is intrinsic and implies that the estimator of the parameter of interest itself has a distribution. Thus Fisher proved himself to be a dualist in making the explicit distinction between sample and population. Neither Big Data, nor full census research or any other attempt to take into account the whole of reality or a world encoded or encrypted in data, can compensate for it. Although many aspects have remained undiscussed in this contribution we hope to have shown that TMLE/SL contributes to the intended reconciliation between inferential statistics and computational science, and that both, rather than being in contradiction, should be integrating parts in any concept of Data Science.

8 Concluding remark: Targeted Learning and Big Data

The expansion of available data has resulted in a new field often referred to as Big Data. Some advocate that big data changes the perspective on statistics: e.g. since we measure everything, why do we still need statistics? Clearly, big data refers to measuring (possibly, very) high dimensional data on a very large number of units. The truth is that there will never be enough data so that careful design of studies and interpretation of data is not needed anymore.

To start with, lots of bad data is useless, so one will need to respect the experiment that generated the data in order to carefully define the target parameter and its interpretation, and design of experiments is as important as ever so that the target parameters of interest can actually be learned.

Even though the standard error of a simple sample mean might be so small that there is no need for confidence intervals, one is often interested in much more complex statistical target parameters. For example, consider the average treatment effect of our running example, which is not a very complex parameter relative to many other parameters of interest such as an optimal individualized treatment rule. Evaluation of the average treatment effect based on a sample (i.e., substitution estimator obtained by plugging in the empirical distribution of the sample) would require computing the mean outcome for each possible strata of treatment and covariates. Even with $n = 10^{12}$ observations, most of these strata will be empty for reasonable dimensions of the covariates, so that this pure empirical estimator is not defined. As a consequence, we will need smoothing (i.e. s super learning), but really, we will also need targeted learning for unbiased estimation and valid statistical inference.

Targeted learning was developed in response to high dimensional data, in which reasonably sized parametric models are simply impossible to formulate and are immensely biased anyway. The high dimension of the data only emphasizes the need for realistic (and thereby large semiparameric) models, target parameters defined as features of the data distribution instead of coefficients in these parametric models, and targeted learning.

The massive dimension of the data does make it appealing to not be necessarily restricted by a priori specification of the target parameters of interest so that targeted learning of data adaptive target parameters discussed above is particularly important future area of research providing an important additional flexibility without to give up on statistical inference.

One possible consequence of the building of large data bases that collect data on total populations is that the data might correspond with observing a single process, like a community of individuals over time, in which case one cannot assume that the data is the realization of a collection of independent experiments, the typical assumption most statistical methods rely upon. That is, data cannot be represented as random samples from some target population since we sample all units of the target population. In these cases, it is important to document the connections between the units so that one can pose statistical models that rely on the a variety of conditional independence assumptions (as in causal inference for networks developed in (71). That is, we need targeted learning for dependent data whose data distribution is modeled through realistic conditional independence assumptions.

Such statistical models do not allow for statistical inference based on simple methods such as the bootstrap (i.e., sample size is 1), so that asymptotic theory for estimators based on influence curves and the state of the art advances in weak convergence theory is more crucial than ever. That is, the state of the art in probability theory will only be more important in this new era of Big Data. Specifically, one will need to establish convergence in distribution of standardized estimators in these settings in which the data corresponds with the realization of one gigantic random variable for which the statistical model assumes a lot of structure in terms of conditional independence assumptions.

Of course, targeted learning with Big Data will require the programming of scalable algorithms, putting fundamental constraints on the type of super-learners and TMLE.

Clearly, Big Data does require integration of different disciplines, fully respecting the advances made in the different fields such as computer science, statistics, probability theory, and in scientific knowledge that allows us to reduce the size of the statistical model and to target the relevant target parameters. Funding agencies need to recognize this so that money can be spent in the best possible way: the best possible way is not to give up on theoretical advances, but the theory has to be relevant to address the real challenges that come with real data. The Biggest Mistake we can make in this Big Data Era is to give up on deep statistical and probabilistic reasoning and theory, and corresponding education of our next generations, and somehow think that it is just a matter of applying algorithms to data.

Acknowledgement

The authors thank the reviewers for their very helpful comments which improved the manuscript substantially. This research was supported by an NIH grant 2R01AI074345.

References

[1] L.B. Balzer and M.J. van der Laan. Estimating effects on rare outcomes: Knowledge is power. Technical Report 310, Division of Biostatistics, University of California, Berkeley, 2013.

[2] H. Bang and J.M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–972, 2005.

[3] 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. *Statistics in Medicine*, 28:152–72, 2009.

[4] O. Bembom and M.J. van der Laan. A practical illustration of the importance of realistic individualized treatment rules in causal inference. *Electronic Journal of Statistics*, 2007.

[5] P.J. Bickel, C.A.J. Klaassen, Y. Ritov, and J. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, 1997.

[6] Jordan Brooks, Mark J van der Laan, and Alan S Go. Targeted maximum likelihood estimation for prediction calibration. *International Journal of Biostatistics*, 2012.

[7] Jordan Brooks, Mark J van der Laan, Daniel E Singer, and Alan S Go. Targeted minimum loss-based estimation of causal effects in right-censored survival data with time-dependent covariates: Warfarin, stroke, and death in atrial fibrillation. *Journal of Causal Inference*, 2013.

[8] M. Carone, M. Petersen, and M.J. van der Laan. Targeted minimum loss based estimation of a casual effect using interval censored time to event data. In Karl E. Peace (eds) Ding-Geng (Din) Chen, Jianguo Sun, editor, *Interval Censored Time To Event Data: Methods and Applications*. Chapman & Hall/CRC, New York, 2012.

[9] A. Chambaz, N. Pierre, and M.J. van der Laan. Estimation of a non-parametric variable importance measure of a continuous exposure. Technical Report 292, UC Berkeley, 2013. to appear in Electronic Journal of Applied Statistics.

[10] A. Chambaz and M.J. van der Laan. Targeting the optimal design in randomized clinical trials with binary outcomes and no covariate, simulation study. *Int J Biostat*, 7(1):33–, 2011. Working paper 258,www.bepress.com/ucbbiostat.

[11] A. Chambaz and M.J. van der Laan. Targeting the optimal design in randomized clinical trials with binary outcomes and no covariate, theoretical study. *Int J Biostat*, 7(1):1–32, 2011. Working paper 258, www.bepress.com/ucbbiostat.

[12] Iván Díaz and Mark van der Laan. Super learner based conditional density estimation with application to marginal structural models. *The International Journal of Biostatistics*, 7(1):38, 2011.

[13] Iván Díaz and Mark van der Laan. Population intervention causal effects based on stochastic interventions. *Biometrics*, 68(2):541–549, 2012.

[14] Iván Díaz and Mark J van der Laan. Assessing the causal effect of policies: An example using stochastic interventions. *International Journal of Biostatistics*, 2013, In press.

[15] Iván Díaz and Mark J van der Laan. Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems. *International Journal of Biostatistics*, 2013, In press.

[16] Iván Díaz and Mark J van der Laan. Targeted data adaptive estimation of the causal dose response curve. *Journal of Causal Inference*, 2013, In press.

[17] Ivan Diaz and M.J. van der Laan. Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems. Technical Report 303, Division of Biostatistics, University of California, Berkeley, 2012. Submitted to IJB, also technical report http://www.bepress.com/ucbbiostat/paper303.

[18] R.D. Gill. Non- and semiparametric maximum likelihood estimators and the von Mises method (part 1). *Scand J Stat*, 1989.

[19] R.D. Gill, M.J. van der Laan, and J.A. Wellner. Inefficient estimators of the bivariate survival function for three models. *Annales de l'Institut Henri Poincaré*, 31:545–597, 1995.

[20] 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.

[21] S. Gruber and M.J. van der Laan. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *International Journal of Biostatistics*, 6:article 26, www.bepress.com/ijb/vol6/iss1/26, 2010.

[22] S. Gruber and M.J. van der Laan. Consistent causal effect estimation under dual misspecification and implications for confounder selection procedure. *Statistical Methods in Medical Research*, February 2012.

[23] S. Gruber and M.J. van der Laan. Targeted minimum loss based estimator that outperforms a given estimator. *The International Journal of Biostatistics*, 8(1):Article 11, doi: 10.1515/1557–4679.1332, 2012.

[24] P.W. Holland. Statistics and causal inference. *J Am Stat Assoc*, 81(396):945–960, 1986.

[25] Samuel D Lendle, Bruce Fireman, and Mark J van der Laan. Balancing score adjusted targeted minimum loss-based estimation. 2013.

[26] Samuel D Lendle, Bruce Fireman, and Mark J van der Laan. Targeted maximum likelihood estimation in safety analysis. *Journal of clinical epidemiology*, 66(8):S91–S98, 2013.

[27] Samuel D Lendle, Meenakshi S Subbaraman, and Mark J van der Laan. Identification and efficient estimation of the natural direct effect among the untreated. *Biometrics*, pages 1–8, 2013.

[28] M.J. van der Laan R. Platt M. Klei3 M. Schnitzer, E. Moodie. Targeted minimum loss based estimator that outperforms a given estimator. *to appear in Biometrics*, page http://biostats.bepress.com/ucbbiostat/paper304, 2013.

[29] 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, *in Design, Summarization, Analysis & Interpretation of Clinical Trials with Time-to-Event Endpoints*. Chapman and Hall, 2009.

[30] 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, 2009.

[31] 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, 2009.

[32] R. Neugebauer, J.A. Schmittdiel, and M.J. van der Laan. Targeted learning in real-world comparative effectiveness research with time-varying interventions. Technical Report No. HHSA29020050016I, The Agency for Healthcare Research and Quality, 2013. http://diabetestranslation.org/en/news$_publications/Reports$.

[33] R. Neugebauer, M.J. Silverberg, and M.J. van der Laan. Observational study and individualized antiretroviral therapy initiation rules for reducing cancer incidence in HIV-infected patients. 272, Division of Biostatistics, University of California, Berkeley, 2010.

[34] J. Neyman. On the application of probability theory to agricultural experiments. *Statistical Science*, 5:465–480, 1990.

[35] J. Pearl. *Causality: models, reasoning, and inference.* Cambridge, New York, 2nd edition, 2009.

[36] M. Petersen, J. Schwab, S. Gruber, N. Blaser, M. Schomaker, and M.J. van der Laan. Targeted minimum loss based estimation of marginal structural working models. *Journal of Causal Inference*, submitted, technical report http://biostats.bepress.com/ucbbiostat/paper312/, 2013.

[37] M.L. Petersen and M.J. van der Laan. A general roadmap for the estimation of causal effects. Unpublished, Division of Biostatistics, University of California, Berkeley, 2012.

[38] E.C. Polley, Sherri Rose, and M.J. van der Laan. Super learning. In M.J. van der Laan and S. Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Data.* Springer, New York Dordrecht Heidelberg London, 2012.

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

[40] J. M. Robins, A. Rotnitzky, and D. O. Scharfstein. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology, the Environment and Clinical Trials*, IMA Volumes in Mathematics and Its Applications. Springer, 1999.

[41] J.M. Robins. Addendum to: "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect" [Math. Modelling **7** (1986), no. 9-12, 1393–1512; MR 87m:92078]. *Comput. Math. Appl.*, 14(9-12):923–945, 1987.

[42] J.M. Robins. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *J Chron Dis (40, Supplement)*, 2:139s–161s, 1987.

[43] J.M. Robins, L. Li, E. Tchetgen, and A.W. van der Vaart. Higher order influence functions and minimax estimation of non-linear functionals. In *Essays in Honor of David A. Freedman*, IMS, Collections Probability and Statistics, pages 335–421. Springer New York, 2008.

[44] S. Rose, R.J.C.M. Starmans, and M.J. van der Laan. Targeted learning for causality and statistical analysis in medical research. Technical Report 297, Division of Biostatistics, University of California, Berkeley, 2011.

[45] S. Rose and M.J. van der Laan. Simple optimal weighting of cases and controls in case-control studies. *The International Journal of Biostatistics*, page http://www.bepress.com/ijb/vol4/iss1/19/., 2008.

[46] S. Rose and M.J. van der Laan. Why match? [investigating matched case-control study designs with causal effect estimation. *The International Journal of Biostatistics*, page http://www.bepress.com/ijb/vol5/iss1/1/., 2009.

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

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

[49] 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.

[50] 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.

[51] A. Rotnitzky, Q. Lei, M. Sued, and J. M. Robins. Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99(2):439–456, doi: 10.1093/biomet/ass013, 2012.

[52] A. Rotnitzky, D. Scharfstein, S. Ting-Li Su, and J. Robins. Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics*, 57(1):103–113, 2001.

[53] D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ Psychol*, 64:688–701, 1974.

[54] D.B. Rubin. *Matched Sampling for Causal Effects.* Cambridge University Press, Cambridge, MA, 2006.

[55] D.B. Rubin and M.J. van der Laan. Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. *The International Journal of Biostatistics*, Vol. 4, Iss. 1, Article 5, 2008.

[56] S. Sapp, M.J. van der Laan, and K. Page. Targeted estimation of variable importance measures with interval-censored outcomes. Technical Report 307, UC Berkeley, 2013. Submitted to IJB.

[57] D. O. Scharfstein, A. Rotnitzky, and J. M. Robins. Adjusting for nonignorable drop-out using semiparametric non-response models (with discussion). *Journal of the American Statistical Association*, 94:1096–1146, 1999.

[58] D.O. Scharfstein, A. Rotnitzky, and J.M. Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models, (with discussion and rejoinder). *J Am Stat Assoc*, 94:1096–1120 (1121–1146), 1999.

[59] R.J.C.M. Starmans. Models, inference and truth: Probabilistic reasoning in the information era. In M.J. van der Laan and S. Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Studies*, pages 1–20. Springer, New York, 2011.

[60] R.J.C.M. Starmans. Picasso, Hegel and the era of big data (in dutch). *Stator*, 2(24), 2013.

[61] R.J.C.M. Starmans and M.J. van der Laan. Inferential statistics versus machine learning; a prelude to reconciliation (in dutch). *Stator*, 2(24), 2013.

[62] O.M. Stitelman and M.J. van der Laan. Collaborative targeted maximum likelihood for time to event data. Technical Report 260, Division of Biostatistics, University of California, Berkeley, 2010.

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

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

[65] Meenakshi Sabina Subbaraman, Samuel Lendle, Mark Laan, Lee Ann Kaskutas, and Jennifer Ahern. Cravings as a mediator and moderator of drinking outcomes in the combine study. *Addiction*, 2013.

[66] C. Tuglus and M.J. van der Laan. Targeted methods for biomarker discovery, the search for a standard. *UC Berkeley Working Paper Series*, page http://www.bepress.com/ucbbiostat/paper233/., 2008.

[67] C. Tuglus and M.J. van der Laan. Modified FDR controlling procedure for multi-stage analyses. *Stat Appl Genet Mol*, 8(1):Article 12, 2009.

[68] C. Tuglus and M.J. van der Laan. Targeted methods for biomarker discoveries. In M.J. van der Laan and S. Rose, *Targeted Learning:Causal Inference for Observational and Experimental Data*, chapter 22. Springer, New York, 2011.

[69] M.J. van der Laan. Estimation based on case-control designs with known prevalance probability. *The International Journal of Biostatistics*, page http://www.bepress.com/ijb/vol4/iss1/17/, 2008.

[70] M.J. van der Laan. Targeted maximum likelihood based causal inference: Part I. *Int J Biostat*, 6(2):Article 2, 2010.

[71] M.J. van der Laan. Causal inference for networks. Technical Report 300, UC Berkeley, 2012. http://biostats.bepress.com/ucbbiostat/paper300,to appear in Journal of Causal Inference.

[72] M.J. van der Laan. Statistical inference when using data adaptive estimators of nuisance parameters. Technical Report 302, Division of Biostatistics, University of California, Berkeley, submitted to IJB, 2012.

[73] M.J. van der Laan. Targeted learning of an optimal dynamic treatment and statistical inference for its mean outcome. Technical Report 317, UC Berkeley, 2013. http://biostats.bepress.com/ucbbiostat/paper317, to appear in Journal of Causal Inference.

[74] M.J. van der Laan, L.B. Balzer, and M.L. Petersen. *Journal of Statistical Research*, 46(2):113–156.

[75] 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, Division of Biostatistics, University of California, Berkeley, November 2003.

[76] M.J. van der Laan, S. Dudoit, and A.W. van der Vaart. The cross-validated adaptive epsilon-net estimator. *Statistics and Decisions*, 24(3):373–395, 2006.

[77] M.J. van der Laan and S. Gruber. Collaborative double robust penalized targeted maximum likelihood estimation. *The International Journal of Biostatistics*, 2010.

[78] M.J. van der Laan, A.E. Hubbard, and S. Kherad. Statistical inference for data adaptive target parameters. Technical Report 314, UC Berkeley, June 2013. http://biostats.bepress.com/ucbbiostat/paper314, Revised for Biometrics.

[79] M.J. van der Laan and M.L. Petersen. Targeted learning. In *Ensemble Machine Learning*, chapter pages 117–156, ISBN 978-1-4419-9326-7. Springer, New York, 2012.

[80] M.J. van der Laan, E. Polley, and A. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(25), 2007.

[81] M.J. van der Laan and J.M. Robins. *Unified methods for censored longitudinal data and causality*. Springer, New York, 2003.

[82] M.J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

[83] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Emprical Processes*. Springer-Verlag New York, 1996.

[84] A.W. van der Vaart, S. Dudoit, and M.J. van der Laan. Oracle inequalities for multi-fold cross-validation. *Statistics and Decisions*, 24(3):351–371, 2006.

[85] 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.

[86] H. Wang, S. Rose, and M.J. van der Laan. Finding quantitative trait loci genes. In M.J. van der Laan and S. Rose, *Targeted Learning:Causal Inference for Observational and Experimental Data*, chapter 23. Springer, New York, 2011.

[87] W. Zheng, M.L. Petersen, and M.J. van der Laan. Estimating the effect of a community-based intervention with two communities. *Journal of Causal Inference*, 1(Issue 1):83–106, 2013.

[88] W. Zheng and M.J. van der Laan. Cross-validated targeted minimum loss based estimation. In M.J. van der Laan and S. Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Studies*. Springer, New York, 2011.

[89] W. Zheng and M.J. van der Laan. Causal mediation in a survival setting with time-dependent mediators. Technical Report 295, Division of Biostatistics, University of California, Berkeley, 2012.

[90] W. Zheng and M.J. van der Laan. Targeted maximum likelihood estimation of natural direct effects. *International Journal of Biostatistics*, 8(Issue 1), 2012.