Biostatistics (2019) **0**, 0, *pp.* 1–6
doi:10.1093/biostatistics/kxz042

Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning

IVÁN DÍAZ*

Division of Biostatistics, Weill Cornell Medicine, 402 East 67th Street, New York, NY 10065, USA
ild2005@med.cornell.edu

SUMMARY

In recent decades, the fields of statistical and machine learning have seen a revolution in the development of data-adaptive regression methods that have optimal performance under flexible, sometimes minimal, assumptions on the true regression functions. These developments have impacted all areas of applied and theoretical statistics and have allowed data analysts to avoid the biases incurred under the pervasive practice of parametric model misspecification. In this commentary, I discuss issues around the use of data-adaptive regression in estimation of *causal inference* parameters. To ground ideas, I focus on two estimation approaches with roots in semi-parametric estimation theory: targeted minimum loss-based estimation (TMLE; van der Laan and Rubin, 2006) and double/debiased machine learning (DML; Chernozhukov *and others*, 2018). This commentary is not comprehensive, the literature on these topics is rich, and there are many subtleties and developments which I do not address. These two frameworks represent only a small fraction of an increasingly large number of methods for causal inference using machine learning. To my knowledge, they are the only methods grounded in statistical semi-parametric theory that also allow unrestricted use of data-adaptive regression techniques.

Keywords: Causal inference; Double/debiased machine learning; Machine learning; Targeted minimum loss-based estimation.

1. AN IMPORTANT DISTINCTION BETWEEN CAUSAL AND STATISTICAL INFERENCE

I subscribe to the useful notion that causal inference is a two-step process. The first step does not involve data. Instead, one uses subject-matter knowledge to pose a causal model, such as the Neyman–Rubin model (Rubin, 1974), a non-parametric structural equation model (Pearl, 2000), a single-world intervention graph (Richardson and Robins, 2013), or a finest fully randomized causally interpreted structured tree graph (Robins, 1986) (a demonstration of a causal model using directed acyclic graphs is published in this commentary series [Robinson *and others*, 2019]). The goal of this causal model is to test whether the causal effect, defined in terms of latent counterfactual variables, is identifiable. Identifiability is the task of representing the causal effect as a function of the probability distribution of the observed data, under a set of untestable assumptions. I refer to this function, which often depends on the probability

*To whom correspondence should be addressed.

distribution of the data only through certain nuisance parameters (e.g., the outcome regression or the propensity score), as the estimand. In the second step, one resorts to data analysis techniques to obtain an estimate of the estimand. TMLE and DML are useful only in the second step, they are methods to estimate features of an observed data distribution, and do not require an underlying causal model or interpretation.

The above point leads me to highlight an important distinction between the concept of identification bias, which measures the difference between the estimand and the causal effect, and the concept of estimation bias, which measures the difference between an estimate obtained from data and the estimand. Causal bias can only be attenuated with better causal models; estimation bias can only be addressed through better estimation methods. An emblematic example of the problems that arise when this distinction is not clear is the widespread but wrong belief that predictive variable selection can be used for confounder selection. This problem that can only be solved through the use of a causal model encoding subject-matter knowledge not present in the data (Hernán, 2019). Predictive variable selection methods are ill-suited to address identification biases, such as the bias that appears when conditioning on a collider of predictors of the outcome and treatment which are not confounders (M-bias, Shrier, 2008). When performing variable selection for causal inference, it is important to distinguish between two goals: variable selection for identification and variable selection for estimator performance. Data-adaptive methods for the former goal (e.g., Häggström, 2018) invariably require a causal model encoding some auxiliary subject-matter knowledge not present in the data. Data-adaptive methods for the latter goal (e.g., collaborative TMLE, (Zheng and van der Laan, 2011) outcome-adaptive lasso, Shortreed and Ertefaie, 2017, etc.) aim to improve the performance of the estimator with respect to the target estimand (i.e., reduce mean squared error), and do not solve problems related to identification bias.

The error in the opposite direction is just as common: researchers often spend a great deal of effort positing the best possible causal model to attenuate identification bias, only to squander those efforts in the estimation phase, using biased estimation techniques such as the parametric g-formula. Even though the identification and estimation stages should generally be kept separate, there is sometimes a useful feedback between them. For example, in randomized trials, it is well known that there are at least two identifiability results, corresponding to unadjusted and adjusted estimands. If pre-treatment covariates are predictive of the outcome, the adjusted estimand is known to have a smaller efficiency bound than the unadjusted one. The adjusted estimand should thus be preferred as it allows more precise estimation of the causal effect. This is an example in which properties of the estimand may be used to inform on the optimal identification strategy. In the next section, I will address ways in which data-adaptive machine learning methods can be used to address estimation bias in general observational studies.

2. HOW TMLE AND DML BRIDGE A GAP BETWEEN MACHINE LEARNING AND STATISTICAL SCIENCE

A foundational requirement in statistical science is the ability to quantify the uncertainty associated with an estimate. This is one of the reasons why parametric models are popular, even though they are widely accepted to produce biased estimates. When using parametric methods, classical tools such as the central limit theorem and the law of large numbers may be used to obtain convenient asymptotic approximations to the sampling distribution of the estimators, readily yielding formulas for computing uncertainty measures such as confidence intervals. Despite this convenience, the formulas derived are defective under model misspecification: they correctly quantify the uncertainty around an incorrect target value. An alarming consequence is that all confidence intervals based on misspecified parametric models contain the target estimand with probability that converges to zero as sample size grows. Flexible machine learning regression may be used to alleviate this bias. However, novel analytical tools (e.g., empirical process theory) are required to characterize and address the bias of the estimators, and to derive sampling distributions and quantify uncertainty around the estimates.

The first of DML and TMLE to appear in the literature is the targeted minimum loss-based estimator. TMLE can be described as a method to construct plug-in estimators of a target estimand. While a naive plug-in estimator would proceed by plugging in machine learning estimates of the nuisance quantities in the estimand formula, TMLE adds an additional step that "forces" the nuisance estimates to solve a set of user-defined estimating equations. These equations are derived from the mathematical study of the estimand, and are useful to endow the plug-in estimator with desirable properties. For instance, the original TMLE formulation was inspired by an earlier observation that naive plug-in estimators incorrectly trade-off bias and variance, accepting more bias than necessary (Koshevnik and Ya Levit, 1977). For a class of target estimands (called pathwise differentiable) the first-order bias can be described as the expectation of a function of the observed data and the nuisance parameters (this function is called the canonical gradient, or efficient influence function). The insight behind the original TMLE formulation was to construct machine learning estimators of the nuisance parameters with the additional property that the empirical version of the first-order bias is set to zero. The upshot is that the resulting plug-in estimator is approximately unbiased. This insight resulted in the proposal of an algorithm in which initial data-adaptive estimators of the nuisance parameters are iteratively tilted towards a solution of this first-order de-biasing equation. The generalization of this framework, in which initial machine learning estimators of nuisance quantities are tilted towards solutions of general estimating equations, has proven successful in solving important estimation problems. For example, it was recently discovered that tilting machine learning estimates towards solutions of certain estimating equations related to the second-order bias term is useful to endow the TMLE with a doubly robust Gaussian asymptotic distribution (Benkeser *and others*, 2017), a problem whose solution remained elusive for many years.

A limitation of the original TMLE formulation is that it imposed an assumption on the statistical model for the nuisance parameters. The "uniform" central limit theorem necessary to derive the asymptotic distribution for the TMLE requires that the nuisance functions are elements of Donsker classes, which are classes of functions with limited entropy. This assumption is problematic in the context of machine learning because it is not satisfied by many estimation methods, e.g., those which yield functions with unbounded variation. The solution to this problem, named cross-validated TMLE (Zheng and van der Laan, 2011), uses sample-splitting to obtain out-of-sample estimates of the nuisance parameters, an idea that dates back to at least (Bickel, 1982). Because the nuisance estimates are constructed using out-of-sample data, asymptotic arguments for the validation sample can be made conditional on the training sample, and then averaged across validation samples, thus avoiding the Donsker condition.

Double/debiased machine learning is a more recent development that also allows the use of machine learning estimates of nuisance quantities. Like TMLE, DML is motivated by the fact that a naive estimator based on machine learning yields a non-negligible first-order bias. Also like TMLE, DML avoids the Donsker condition by using out-of-sample estimates of the nuisance quantities. However, DML operates in a slightly different framework, where the estimand is characterized as the solution to a population score equation. In DML, first-order bias is defined with respect to a property called Neyman orthogonality, which ensures that the score equations that define the estimand are first-order invariant to small changes in the nuisance parameters. The use of Neyman-orthogonal score equations is thus proposed as a means to remove the first-order bias. Because the estimand is defined as the solution to a score equation, the DML framework is well suited for traditional semi-parametric models indexed by one finite- and one infinite-dimensional component, where the finite-dimensional component is the target of inference. When the target of inference is explicitly defined as a functional of the data distribution and the parameter is pathwise differentiable, such as the average treatment effect (ATE), the efficient influence function is Neyman orthogonal. Implementation of DML estimators in these cases amounts to simply computing a cross-fitted (i.e., prediction is out-of-sample) version of the well-known one-step correction (Pfanzagl, 1982) (e.g., computing a cross-fitted augmented inverse-probability weighted estimator). Compared to plug-in estimators like the TMLE, solutions to estimating equations have the drawback that they may yield estimates outside the estimand parameter space. Furthermore, it has been noted that some important

problems such as that of endowing estimators with doubly robust asymptotic distributions cannot be solved within an estimating equation framework (Benkeser *and others*, 2017).

Both the TMLE and DML enjoy optimal statistical properties such as $n^{1/2}$-rate asymptotic normality and efficiency under consistency assumptions on the nuisance estimators. In the case of the ATE, a sufficient assumption is that both the outcome regression and the propensity score are consistent at a rate faster than $n^{1/4}$, much slower than the optimal $n^{1/2}$-rate achieved by correctly specified parametric models. The $n^{1/4}$-rate is achievable by many data-adaptive regression methods, such as L1 penalization of generalized linear models (Bickel *and others*, 2009), random forests (Wager and Walther, 2015), neural networks (Chen and White, 1999), and the highly adaptive lasso (Benkeser and van der Laan, 2016). Because it is hard to know *a priori* which method will yield the best performance for a given dataset, it is often recommended to use cross-validation estimator selection or a cross-validation ensemble of estimators such as the super learner (van der Laan *and others*, 2007).

DML and TMLE estimate parameters that are smooth in the sense that they admit a first-order representation, which allows a characterization of the bias and efficiency bound and the development of parametric-rate asymptotically Gaussian estimators. Many parameters motivated by causal inference, such as the conditional average treatment effect, the dose–response curve for a continuous treatment, and the optimal treatment rule are fundamentally harder to estimate nonparametrically because they do not admit a suitable first-order representation. Data-adaptive methods have been developed to estimate these parameters; examples are generalized random forests (Wager and Athey, 2018), doubly robust pseudo-outcome kernel regression (Kennedy *and others*, 2017; Wager and Athey, 2018), outcome weighted learning (Zhao *and others*, 2012), Q-learning (Goldberg and Kosorok, 2012), and ensemble learners (Díaz *and others*, 2018). Asymptotic results that allow the computation of formal p-values and confidence intervals exist for some of these methods, though the convergence rate is often not parametric. For some other methods, the only available theoretical guarantees are in the form of risk bounds and oracle inequalities, and the solution to quantifying the uncertainty around the estimates remains an open problem.

3. Conclusion and some thoughts

One of the pillars of statistical science, which sets it apart from other data analysis fields, is the ability to perform formal statistical inference. TMLE and DML are examples of methods that successfully bridge the gap between machine learning and statistical science, allowing data analysts to use machine learning while obtaining valid methods for formal statistical inference. TMLE and DML achieve this by embracing the technical challenges that come with realistic statistical models, instead of shunning them in favor of the analytical convenience of unscientific parametric methods. In a big data world dominated by complex causal inference questions and machine learning applications, the continued success of statistics as a profession depends on its ability to adapt to these new tools and respond to real-life scientific problems while being truthful to its foundational principles. Formulating and answering meaningful scientific problems requires causal inference thinking. Incorporating machine learning into our estimation toolbox while performing statistical inference requires the integration of empirical processes, high-dimensional statistics, and semi- and non-parametric statistics. I believe these points should be kept in mind as our statistics discipline goes through the changes brought about by the advent of data science, particularly as these changes relate to PhD and MS curricula and data analysis quality standards for applied research.

Acknowledgments

Conflict of Interest: None declared.

References

Benkeser, D. *and others* (2017). Doubly robust nonparametric inference on the average treatment effect. *Biometrika* **104**, 863–880.

Benkeser, D. and van der Laan, M. (2016). The highly adaptive lasso estimator. In: *Proceedings of the International Conference on Data Science and Advanced Analytics. IEEE International Conference on Data Science and Advanced Analytics*. pp. 689–696.

Bickel, P. J. (1982). On adaptive estimation. *The Annals of Statistics* **10**, 647–671.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* **37**, 1705–1732.

Chen, X. and White, H. (1999). Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory* **45**, 682–691.

Chernozhukov, V. *and others* (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* **21**, C1–C68.

Díaz, I., Savenkov, O. and Ballman, K. (2018). Targeted learning ensembles for optimal individualized treatment rules with time-to-event outcomes. *Biometrika* **105**, 723–738.

Goldberg, Y. and Kosorok, M. R. (2012). Q-learning with censored data. *Annals of Statistics* **40**, 529.

Häggström, J. (2018). Data-driven confounder selection via Markov and Bayesian networks. *Biometrics* **74**, 389–398.

Hernán, M. A. (2019). Comment: Spherical cows in a vacuum: data analysis competitions for causal inference. *Statistical Science* **34**, 69–71.

Kennedy, E. H. *and others* (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* **79**, 1229–1245.

Koshevnik, Y. A. and Ya Levit, B. (1977). On a non-parametric analogue of the information matrix. *Theory of Probability and Its Applications* **21**, 738–753

Pearl J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Pfanzagl, J. (1982). *Contributions to a General Asymptotic Statistical Theory*. Lecture Notes in Statistics.

Richardson, T. S. and Robins, J. M. (2013). Single world intervention graphs (SWIGs): a unification of the counterfactual and graphical approaches to causality. *Working Paper*, 128(30). Center for the Statistics and the Social Sciences, University of Washington Series, pp 2013.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling* **7**, 1393–1512.

Robinson, W. R., Renson, A. and Naimi, A. I. (2019). Teaching yourself about structural racism will improve your machine learning. *Biostatistics*. doi:10.1093/biostatistics/kxz040.

rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.

Shortreed, S. M. and Ertefaie, A. (2017). Outcome-adaptive lasso: variable selection for causal inference. *Biometrics* **73**, 1111–1122.

shrier, I. (2008). Letter to the Editor. *Statistics in Medicine* **27**, 2740–2741.

van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics* **2**, 1–40.

van der Laan, M. J., Polley, E. C. and Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology* **6**, 1.

WAGER, S. AND ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* **113**, 1228–1242.

WAGER, S. AND WALTHER, G. (2015). Adaptive concentration of regression trees, with application to random forests. arXiv [math.ST]. http://arxiv.org/abs/1503.06388.

ZHAO, Y. *and others* (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* **107**, 1106–1118.

ZHENG, W. AND VAN DER LAAN, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In: *Targeted Learning*. New York, NY: Springer, pp. 459–474.

[*Received September 25, 2019; revised September 25, 2019; accepted for publication September 25, 2019*]