Statistical Science
2019, Vol. 34, No. 1, 72–76
https://doi.org/10.1214/18-STS680
Main article: https://doi.org/10.1214/18-STS667

Comment: Will Competition-Winning Methods for Causal Inference Also Succeed in Practice?

Qingyuan Zhao, Luke J. Keele and Dylan S. Small

Abstract. First, we would like to congratulate the authors for successfully hosting the causal inference data competition (referred to as Competition henceforth) and contributing a unique and thought-provoking article to the literature. The authors have provided a comprehensive and timely platform to evaluate the ever-growing number of methods used for covariate adjustment in observational studies. In our comment, we don't generally question the results of the competition, but we do wish to emphasize several other key elements about the role statistics plays in causal inference and observational studies.

Key words and phrases: Observational studies, machine learning, study design.

TESTING GROUNDS FOR CAUSAL INFERENCE: "IN VITRO" VERSUS "IN VIVO"

One of the main conclusions learned from this contest was that, "methods that flexibly model the response surface perform better overall than methods that fail to do so." In view of Breiman's (2001) famous dichotomy, this would appear to be another triumph for the algorithmic culture of statistical modeling. Just like hundreds of online machine learning competitions (for example those hosted by *Kaggle*), highly adaptive black box algorithms are shown once again to outperform "traditional" statistical methods such as linear regression.

However, unlike the winners of machine learning competitions, we believe it is not obvious that a "competition-winning" method for causal inference should be immediately deployed in practice, even if the competition is as comprehensive as the one in the paper being discussed. The first reason for our caution is the inherent differences between predictive inference and causal inference. In predictive tasks, it is often straightforward to evaluate the performance of machine learning methods in real life by simply holding out a test dataset. Thus, it is easy to create a fair testing ground for predictive methods. Unfortunately, causal inference methods cannot be evaluated this way because a successful method needs to predict the outcome in different interventional settings, which are not available in observational datasets.

Besides the simulation-based comparisons, another testing ground for causal inference methods is the within-study comparison, where the control group of a randomized experiment is replaced with an observational comparison group (this is discussed in Section 2.1 of the main paper). A good analogue is *in vitro* (meaning "in the glass" or "in the test tube") versus *in vivo* (meaning "in the living") experiments in biology. How about change the last two sentences to: Like *in vitro* experiments, simulation-based comparisons can test the performance of statistical methods in highly

Qingyuan Zhao is Postdoctoral fellow, Department of Statistics, The Wharton School, University of Pennsylvania, 400 Huntsman Hall, 3730 Walnut Street, Philadelphia, Pennsylvania 19104, USA (e-mail: qyzhao@wharton.upenn.edu). Luke J. Keele is Associate Professor of Applied Statistics, Hospital of the University of Pennsylvania, Silverstein 4, 3400 Spruce Street Philadelphia, Pennsylvania 19104, USA (e-mail: luke.keele@gmail.com). Dylan S. Small is Professor of Statistics, Department of Statistics, Wharton School, University of Pennsylvania, 400 Huntsman Hall, 3730 Walnut Street, Philadelphia, Pennsylvania 19104, USA (e-mail: dsmall@wharton.upenn.edu).

controlled settings, thus they are simple to implement and automate. However, they often make the simplifying assumptions that there is no unmeasured confounding and the overlap assumption is reasonably satisfied, and thus may not reflect the performance in real studies. Like *in vivo* experiments, within-study comparisons yield critical information about the statistical methods in actual practice, but they are often costly and difficult to conduct. The authors should be congratulated for a clear advance in the design of "in vitro" experiments for causal inference methods. Just as the predictive competitions have been so instrumental in advancing the field of machine learning, the competition presented in the article has shed new light on the efficacy of flexible black box methods in causal inference. However, we think further developments of "in vivo" studies are also extremely important to understand the efficacy and limitations of causal inference methods in realistic scenarios. The studies in Cook, Shadish and Wong (2008) and Shadish, Clark and Steiner (2008) are all too rare examples of what can be learned from "in vivo" investigations.

SOURCES OF ESTIMATION ERROR

A key limitation of the "in vitro" design is the lack of consideration of hidden bias in real studies. In general, we can describe the estimation error of any causal effect estimator with the following equation:

$$
\begin{aligned}
&\text{Estimator} - \text{True causal effect} \\
&\quad = \underbrace{\text{Hidden bias}}_{\text{Due to unmeasured confounding}} \\
&\qquad + \underbrace{\text{Misspecification bias}}_{\text{Due to parametric modeling}} \\
&\qquad + \underbrace{\text{Noise}}_{\text{Due to finite sample}} .
\end{aligned} \tag{1}
$$

The first term "hidden bias" is due to poor design of the study and can include unmeasured confounding bias or collider bias. The next two terms reflect the familiar bias-variance tradeoff of statistical estimators.

Before commenting on the competition results, we want to emphasize one point about the decomposition (1). An implicit claim in (1) is that the hidden bias does not depend on the particular statistical method used to analyze the data. In other words, the hidden bias is decided once we determine what data will be collected. A simple conceptual proof of this is to imagine two estimators that both converge to the true causal effect when there is *no* unmeasured confounding. They must also converge to the same limit when there *is* unmeasured confounding, because otherwise the ignorability assumption would be testable by empirical data.

In the competition (and usually any "in vitro" comparison), the hidden bias was fixed at zero. Thus, the submitted methods were judged entirely by their ability to find an appealing bias-variance tradeoff between the last two terms of (1) in a wide range of simulation settings. While such a comparison certainly provides valuable information, we worry that readers will lose the big picture and simply interpret the competition results as saying that using flexible machine learning methods is foolproof for valid causal inference. In actual studies we have been involved with, however, the foremost concern has almost always been the hidden bias, the first term in (1). The rationale is that, while misspecification error is a concern, it is still fixable by statistical methods (at least in principle). as demonstrated by the competition In contrast, once hidden bias is present, it cannot be detected or corrected by any statistical method and will stick with any subsequent analysis. We will discuss more about designing an observational study in the next section.

The usefulness of any "in vitro" comparison thus depends on the relative magnitude of the non-statistical hidden bias and the statistical error, that is, the ratio between the first term and the last two terms in (1). We think it is quite possible that the data generating processes in the competition overstated the magnitude of misspecification bias relative to what is present in most applications. That is, the competition demonstrated clearly that flexible black boxes are very good at minimizing misspecification bias, but how large is this quantity in real data? In a recent paper, Keele and Small (2018) find that in a variety of applications, differences in causal effect estimates between different methods due to misspecification error tended to be quite small. This suggests that while methods flexibly modeling the response surface are more robust when misspecification bias is very large, in many data applications this bias might be much smaller.

DESIGN TRUMPS ANALYSIS: TWO NEW INTERPRETATIONS

In an influential article, Rubin (2008) advocated the motto "design trumps analysis" and argued that objective observational studies must "be carefully designed to approximate randomized experiments, in particular, without examining any final outcome data." Although the three authors of this commentary have different views on how to use outcome data in the design of observational studies, we all agree that Rubin's emphasis on design is appropriate in a broader sense. Here, we want to offer two new interpretations of "design trumps analysis."

Our first interpretation is motivated by the decomposition (1), as the hidden bias due to poor design cannot be corrected by any statistical estimator. We believe the most critical stage of an observational study remains the design, in particular, the selection of the identification strategy. The competition was conducted in a setting where "selection on observables" holds. Under this identification strategy, the quality of the observational study largely depends on which confounders the investigator decides to collect in the design stage of the study. As a side remark, frequently it will be more fruitful to find an alternative identification strategy based on an instrument, a regression discontinuity design, or a natural experiment. We would argue that evidence for a causal effect is strengthened by finding that different plausible identification strategies with different sources of bias yield similar conclusions (Rosenbaum, 2001).

Even when an observational study is based on selection on observables, other aspects of the design stage may reduce hidden bias or increase the quality of the evidence. Specifically, observational studies based on selection on observables will tend to yield better evidence when combined with the use of quasi-experimental devices such as multiple control groups and baseline outcomes that examine whether certain sources of bias are large enough in magnitude to change the qualitative conclusions of a study (Cook, Campbell and Shadish, 2002), the selection of more focused comparisons that reduce unmeasured confounding (Rosenbaum, 2005), making use of parallel treatments (Rosenbaum, 2006), exploiting instances instances of "refutability" including testing for hidden bias using negative control outcomes that aren't thought to be affected by the treatment (Angrist and Krueger, 1999, Rosenbaum, 2002, Lipsitch, Tchetgen Tchetgen and Cohen, 2010), reporting results from sensitivity analyses (Rosenbaum, 1987, Imbens, 2003), and using "pattern specificity" to make claims convincing (Hill, 1965).

The second new interpretation comes from the main finding of the paper being discussed: the competition-winning methods all use flexible models for the response surface and thus have small misspecification bias. Among the methods that do not model the response surface, ones which flexibly model the assignment mechanism are also more robust than those which do not. Furthermore, in Section 7.3 the authors find that as long as the response surface is modeled flexibly, no other characteristic of the methods seems to be associated with the cross-method performance variation. The authors' results thus all point to the same conclusion: the use of a nonparametric model for the response surface (and also the treatment assignment) is more important than the specific nonparametric estimator used.

THE ROLE OF STATISTICS IN CAUSAL INFERENCE

When causal inference methods are applied to answer questions in scientific research, most of the time the investigation team will also include at least one if not several substantive experts. As statisticians, our job is not just to develop methods that are most efficient and robust in the statistical sense. Another important part of our job is to communicate and interact with our collaborators. For this we would like to offer an quote from Box (1979) in his Presidential Address to the American Statistical Association:

> It is widely recognized that the advancement of learning does not proceed by conjecture alone, nor by observation alone, but by an iteration involving both. Certainly, scientific investigation proceeds by such iteration. Examination of empirical data inspires a tentative explanation which, when further exposed to reality, may lead to its modification. This modified explanation is again put in jeopardy by further exposure to reality, and so on, in a continued alternation between induction and deduction.

When collaborating with scientists as described by Box, we can think of at least three other practical concerns beyond the efficiency and robustness properties examined by the Competition:

Exploratory data analysis (EDA): Can meaningful EDA be performed to detect/remove anomalies, visualize the data, and assess assumptions of the statistical inference?

Ease to explain: Is it easy for us to explain the statistical method to our collaborators who may lack the technical skill? Is it easy for our collaborators to explain the method to their peers?

Substantive Input: Can we effectively interact with our collaborators to incorporate their expert knowledge to improve the analysis? Transparency of the analysis may aid such interactions.

In our experience, these are all critical in a successful scientific collaboration. For example, in the context of observational studies, the statistical method should facilitate the removal of treated units that are far away from the support of the control group. This can often be examined after matching by the covariates. As another example, if certain covariates are believed to be more important confounders by our collaborators, our statistical method should be able to incorporate this information. This is straightforward for methods based on the treatment assignment such as matching (cf. Zubizarreta, 2012, Pimentel et al., 2015) and propensity score weighting (cf. Zhao, 2019) by requiring more stringent balancing constraint on these covariates. These methods may perform slightly (or even much) worse in the competition because the response surface is not explicitly modeled, but their transparency may better aid the data visualization and other collaborative needs listed above.

Unfortunately, it is nearly impossible to incorporate these considerations into simulation-based comparisons. In fact, such "in vitro" testing is more challenging for design-based method because domain knowledge cannot be used (the covariates are usually coded as `X1,X2,...,` with no physical meaning attached). In an "in vivo" comparison using five empirical applications, Keele and Small (2018) find that carefully designed matching methods and black box machine learning methods only modeling the regression surface mostly produce identical results. Interestingly, in one case they find that prioritizing certain covariates in matching can substantially change the causal effect estimate. Thus, a priori knowledge can possibly play an important role in observational studies in practice.

CONCLUSION

We want to thank the authors again for their efforts in creating this data competition. The main message that the response surface should be flexibly modeled is well received. We welcome the usage of machine learning in causal inference and we are also thinking about incorporating it in future research designs. One possibility is to use matching with covariate prioritization and black box methods in conjunction and check if the results agree, see Keele and Small (2018).

Our main conclusion is that, while machine learning methods have become indispensable instruments for statisticians in modern large-scale problems, we should not be complacent and become the "data analyst" of the study. On the contrary, we should be even more conscientious about the design of an observational study and continue to find ways to better interact with our scientific collaborators. For this, we would like to end with another quote from Box (1979):

> Please can Data Analysts get themselves together again and become whole Statisticians before it is too late? Before they, their employers, and their clients forget the other equally important parts of the job statisticians should be doing, such as designing investigations and building models? By invention of the concept of Experimental Design, Fisher promoted the statistician from a curator of dusty relics to a valued member of a scientific team, responsible for planning and taking part in the conduct of an investigation. Let us not allow him to be relegated to his previous passive and inferior role by an injudicious choice of a name, "Our Data Analyst" is too close for my liking to "Our Tame Statistician," a poor thing if that is all he is.

REFERENCES

ANGRIST, J. D. and KRUEGER, A. B. (1999). Empirical strategies in labor economics. In *Handbook of Labor Economics* (O. Ashenfelter and D. Card, eds.) **3A** 1277–1366. Elsevier, Amsterdam.

BOX, G. E. (1979). Some problems of statistics and everyday life. *J. Amer. Statist. Assoc.* **74** 1–4.

BREIMAN, L. (2001). Statistical modeling: The two cultures. *Statist. Sci.* **16** 199–231. MR1874152

COOK, T. D., CAMPBELL, D. T. and SHADISH, W. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin, Boston, MA.

COOK, T. D., SHADISH, W. R. and WONG, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *J. Policy Anal. Manage.* **27** 724–750.

HILL, A. B. (1965). The environment and disease: Association or causation? *J. R. Soc. Med.* **58** 295–300.

IMBENS, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *Am. Econ. Rev. Pap. Proc.* **93** 126–132.

KEELE, L. and SMALL, D. (2018). Comparing covariate prioritization via matching to machine learning methods for causal inference using five empirical applications. Preprint. Available at arXiv:1805.03743.

LIPSITCH, M., TCHETGEN TCHETGEN, E. and COHEN, T. (2010). Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology* **21** 383–388.

PIMENTEL, S. D., KELZ, R. R., SILBER, J. H. and ROSENBAUM, P. R. (2015). Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *J. Amer. Statist. Assoc.* **110** 515–527. MR3367244

ROSENBAUM, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* **74** 13–26. MR0885915

ROSENBAUM, P. R. (2001). Replicating effects and biases. *Amer. Statist.* **55** 223–227. MR1963397

ROSENBAUM, P. R. (2002). *Observational Studies*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR1899138

ROSENBAUM, P. R. (2005). Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *Amer. Statist.* **59** 147–152. MR2133562

ROSENBAUM, P. R. (2006). Differential effects and generic biases in observational studies. *Biometrika* **93** 573–586. MR2261443

RUBIN, D. B. (2008). For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* **2** 808–804. MR2516795

SHADISH, W. R., CLARK, M. H. and STEINER, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *J. Amer. Statist. Assoc.* **103** 1334–1343. MR2655714

ZHAO, Q. (2019). Covariate balancing propensity score by tailored loss functions. *Ann. Statist.* **47** 965–993. MR3909957

ZUBIZARRETA, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J. Amer. Statist. Assoc.* **107** 1360–1371. MR3036400