Statistical Science
2019, Vol. 34, No. 1, 94–99
https://doi.org/10.1214/18-STS688
Main article: https://doi.org/10.1214/18-STS667

Rejoinder: Response to Discussions and a Look Ahead

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott and Dan Cervone

Abstract. Response to discussion of Dorie (2017), in which the authors of that piece express their gratitude to the discussants, rebut some specific criticisms, and argue that the limitations of the 2016 Atlantic Causal Inference Competition represent an exciting opportunity for future competitions in a similar mold.

Key words and phrases: Causal inference, competition, machine learning, automated algorithms, evaluation.

1. INTRODUCTION

We are extremely appreciative of the scholars who acted as discussants for our paper. One of the primary goals of creating this competition was to initiate a broader conversation about how we as a community can best evaluate the relative performance of causal inference approaches. This discussion provides an important first step in that direction. In this rejoinder, we respond to each discussion in turn and then conclude with some final thoughts.

2. HERNAN DISCUSSION

Hernan expresses gratitude to the authors for creating and implementing the 2016 Atlantic Causal Inference Conference Competition and then provides a thoughtful discussion of the limitations of this single competition. We agree with many of his points; our response highlights points of agreement and of departure.

2.1 Incorporating Subject-Matter Expertise

Hernan points to the importance of subject matter expertise in causal inference. As an example expert knowledge could be helpful in identifying biasing covariates. We completely agree that there are many settings when such expertise can play a critical role in design and analysis. However it is not clear to us why such knowledge couldn't be incorporated into any of the high-performing techniques? Said another way, we were not intending the results of the competition as a recommendation to ignore or fail to obtain this type of critical information about whatever research question is being addressed.

We do agree that incorporating subject matter expertise into a competition format would be a formidable challenge (particularly since it is sometimes wrong)! However we do not rule out the possibility that a clever design of this sort may emerge in the future. As a case in point, the Karavani et al. discussion below provides an interesting proposal along these lines.

2.2 Future Competitions

Hernan made helpful suggestions for future directions of such competitions. In particular, he advocated for competitions that introduce time-varying treatments and confounders or failure time outcomes. We would welcome the development of such challenges and agree that a different class of estimation methods would be in play in such settings.

Vincent Dorie is an Associate Research Scientist, Columbia University, 475 Riverside Drive, New York, New York 10027, USA (e-mail: vdorie@gmail.com). Jennifer Hill is a Professor of Applied Statistics and Data Science, New York University, 246 Greene Street, 3rd Floor, New York, New York 10003, USA (e-mail: jennifer.hill@nyu.edu). Uri Shalit is an Assistant Professor, Technion, Technion—Israel Institute of Technology, Technion City, Haifa 3200003, Israel (e-mail: urishalit@technion.ac.il). Marc Scott is a Professor of Applied Statistics, New York University, New York University, 246 Greene Street, 3rd Floor, New York, New York 10003, USA (e-mail: marc.scott@nyu.edu). Dan Cervone is Director of Quantitative Research, Los Angeles Dodgers, Dodger Stadium, 1000 Vin Scully Avenue, Los Angeles, California 90012, USA (e-mail: dcervone@gmail.com).

Hernan seemed doubtful that future competitions could incorporate violations of key assumptions such as overlap and ignorability. However, we believe that it is possible to extend the competition framework in ways that do allow for exploration of violations of these ideal conditions. For example, competition organizers could create data where overlap is violated and then participants could decide which estimand they can reliably target. Loss functions could be imposed when evaluating performance to create tradeoffs between simplifying the estimand and the resulting bias or coverage properties. We can also imagine competitions that allow for violations of ignorability but allow for sensitivity analysis or methods that create bounds that allow for the impact of hidden bias. Evaluation criteria would be more difficult to construct in such a setting but it would be a fascinating exercise.

2.3 Performance Under Ideal Conditions

One of the primary tenets of Hernan's discussion, echoed by several of the other discussants, can be simplistically summarized by his statement that "causal inference competitions may only provide advice to practitioners under ideal conditions." Our first response to this is that it is possible to successfully broaden the scope beyond "ideal conditions" as discussed above.

An equally important point is that even under the ideal conditions provided (e.g., i.i.d. data with ignorability and common support satisfied) several widely used methods performed quite poorly across the full range of simulation settings both in the DIY competition and the full Black Box competition. If a method does not perform well under *ideal* circumstances the statistics community might want to think twice about recommending it to our students and colleagues. To reuse the author's provocative image, if we cannot optimize milk production *even with* a gravity-free spherical cow in a vacuum then we need to think harder about what we are recommending as our milking equipment!

3. ZHAO, KEELE, AND SMALL DISCUSSION

Zhao et al. (ZKS) were also generous in their praise of our efforts. However they also expressed concern that our paper might be unduly biased towards automated algorithms and pointed out that it is not clear that contest winning methods "should be immediately deployed in practice." We are sympathetic to this response and would like to understand more deeply what the researchers feel *would be* sufficient evidence to recommend a method to be "deployed in practice." We now address some of their specific arguments about the limitations of our design.

3.1 Hidden Bias Versus Misspecification Bias

We agree with ZKS that it is likely true that in practice bias due to unmeasured confounding is often larger than bias due to model misspecification. Although this is unlikely to be a minority opinion in our community, it is interesting that the vast majority of research performed in this area still posits the assumption of ignorability (no hidden bias). This disconnect between what we think is the biggest problem and what most people choose to focus their research on represents an interesting cultural phenomenon that warrants furthers exploration and may reflect the rewards and incentives in our field.

Our response to this criticism, however, is similar to our response to a similar critique in Hernan's discussion. First, we would love to see future competitions that focus more on situations that allow for hidden bias! And second it is not clear that we have already "solved" the relatively more minor problem of misspecification bias. Given the poor performance of very popular methods in both the DIY and Black Box competitions, focus on misspecification bias seems to still be warranted. There is a classic, statistical argument to be made for tackling one problem at a time, and the lack of universal success with the problem of misspecification bias should give us pause when considering the challenge of hidden bias.

3.2 Design Trumps Analysis; Meaningful Collaboration Is Critical

ZKS further argue that clever design and meaningful collaboration with substantive researchers with deep knowledge of the research question is more important than the analysis model. As a general rule we completely agree with this sentiment. In fact, we have been dismayed to observe how abandoning these important tenets has led to some cautionary tales in the past years (e.g., Lazer et al., 2014). However, even the most carefully designed study will require some method to estimate the treatment effect. How should the researcher choose that method?

One might in fact argue that our competition reflects the position that researchers who implement clever designs find themselves in—they've been able to eliminate hidden bias through their design but still need to estimate a treatment effect. The results of this competition reveal that if that analysis still requires conditioning on covariates then there is a wide variety of performance in estimators even in this ideal circumstance.

What if the researcher is fortunate enough to have data from a completely randomized design (or similar) that does not require such covariate adjustment? In that case results should be less sensitive to choice of an estimator and the results from this competition are less interesting. However it is difficult to imagine that adoption of any of the high-performing methods would cause *harm* in such a situation.

We also note that even clever designs are not fool proof. Randomized experiments can suffer from missing data and noncompliance. Natural experiments that inspire instrumental variables analysis may lead to severe bias if ignorability or exclusion is violated and the percent of compliers is small. Regression discontinuity can be sensitive to functional form and typically only allows inference for a very limited subset of the population.

In sum, we argue that the methods that performed well in this competition should *augment* rather than *supplant* clever design.

3.3 In Vitro Versus in Vivo

ZKS posit that "a successful method needs to be deployed in different intervention settings." However they then argue in favor of "within-study comparisons" rather than our simulated data structure. They liken the former to "in vivo" experiments and the latter to "in vitro" (i.e., test tube) experiments. While we agree that simulated data often lack clear calibration to data from actual studies we would argue that within-study comparisons (also referred to as constructed observational studies) may also present important disconnects with the research questions of primary interest. In some constructions we don't know if ignorability is satisfied. Different incarnations where ignorability is guaranteed, can typically only investigate interventions implemented in rarefied settings (e.g., convenience samples of university students) for highly specialized interventions, with outcomes measured over short periods of time. Most importantly, all of these designs typically represent only a solitary DGP rather than the range of DGPs represented in our competition. While we see the benefits in such evaluation strategies it is important to understand that important trade-offs exist.

4. JENSEN DISCUSSION

We were grateful for Jensen's appreciation of the enormous effort that went into running and analyzing the data from the competition. His discussion takes a largely positive view of our paper and highlights the value of empirical validation of different methodologies—regardless of what we actually learn from this validation—as a complementary exercise in the progress of many areas within Statistics and Computer Science. Indeed, the competition was inspired by similar practices within the machine learning community, such as benchmarking using standard data sets, and valuing forums such as Kaggle as proving grounds for prediction methods. Beyond the motivating factors discussed in the paper, Jensen points out how useful empirical validation is for assessing performance of computationally complex algorithms that don't lend themselves to theoretical evaluations of their properties. We also agree that even with these types of algorithms the implementation is not always straightforward and would encourage future organizers of such challenges to elicit more information from participants about these design choices (i.e., increase the detail and transparency of the "pre-registration" aspect of the competition).

4.1 Future Competitions

Rather than focus on the limitations in our competition design, Jensen supports the concept of future competitions within the causal inference research community and suggests expanding the scope of empirical validation of causal inference methods to different treatment paradigms. Many complications acknowledged in the paper as features we ignored—such as covariate complexity and ignorability—still solicit use of the same methods used in our competition, even though one would expect degraded performance. Different treatment paradigms such as those Jensen provides (including multiple interventions over time or simulateously, and multiple outcomes over time or simultaneoulsy), however, would invite new methodology into the fray.

Jensen also supports deeper exploration regarding the calibration of competition data to real observational data. This is a natural discussion point, but rather difficult to validate. He suggests using real experimental data that exhausts all potential outcomes which is an intriguing idea for competitions such as this one, however that would still only represent one instance of real data as opposed to the general concept of realistic data we strove for by varying features of the response surfaces. However, we recognize that there are always trade-offs across these approaches and are welcoming of new strategies to better evaluate methods.

4.2 Implications

We appreciate Jensen's reiteration of important takeaways from the competition, namely that response surface nonlinearity and poor alignment seemed to present the greatest challenges to the methods surveyed in the competition. In particular, when considering alignment, his discussion reinforces the need for dependence of the treatment assignment and response surface modeling tasks.

5. GRUBER AND VAN DER LAAN DISCUSSION

In their comment, Gruber and van der Laan (GV) provide an overview of the the competition from the perspective of targeted learning. Their entry in the competition, SL + TMLE, estimates the response surface and treatment assignment mechanism using an ensemble machine learning algorithm. During the competition, decisions made by the contestants, including pre-processing or explanatory analysis were not revealed to us, and GV provide some insights both in how they approach causal inference generally, as well as the specific decisions that they made during this competition.

In essence, GV's choices "prime" the learning algorithm, given the limited information made available to participants. For example, they added squared and dichotomized versions of continuous covariates to the feature set before reducing dimensionality with, for example, the lasso. They also describe one procedure used to "[prescreen covariates] in an attempt to exclude IVs from the response surface and treatment models." We wonder to what extent the authors believe that most if not all "prescreening" can be automated? Nevertheless, unpacking the "black box" and uncovering methodological insights such as these was one goal of the competition, and we are pleased that such discussion has been expanded here.

Given the success of the BART entries, GV focus some of their discussion directly and indirectly on this method. First, they note that it was not included in their original submission. Rather than comment directly as to why they did not include it initially, they provide some interesting reasons why one might not always wish to rely solely on BART. We would like to have had them "unpack" that initial decision more fully, as their approach never relies on one method in isolation. They also note that the organizers, DHSSC, used TMLE in the post-competition analysis to target an initial BART estimate. Given the success of the latter, we are still left wondering why BART was not part of the initial ensemble of learners. After all, a stated and a posteriori proven advantage to ensemble learners is their ability to use, or not use information from each method included in the ensemble. They conclude with a brief discussion of the DHSSC analysis of variation in the performance of different methods using oracle and nonoracle features of the data. We view this meta-learning problem (e.g., what methods, functional form and predictors should one use in a particular case) as crucial to any viable automated, "off the shelf" approach. GV contend that the inability to predict a method's performance on a problem from given characteristics validates their approach to the problem more generally, which is to rely on a library of methods used in an adaptive manner. We are a bit concerned that this could become the modern version of a "kitchen sink" approach; their own discussion suggests that it is the interplay between a set of initial decisions and a flexible, theoretically grounded method that is most successful.

6. KARAVANI ET AL. DISCUSSION

Karavani et al. look at the broad question of the present and future of causal inference competitions. They identify several key shortcomings of current simulation-based competitions, such as the one we present in our paper. Beyond many specific problems, the main issue they bring up is the fact that in the full pipeline of an observational study, the choice of inference method is often not necessarily the one with the most impact. Instead, Karavani et al. claim that design questions such as which covariates to include and how to define the cohort are frequently the most crucial, and domain expertise is commonly needed in these steps. This is a point that has been made by some of the other discussants. As a response to these shortcomings, Karavani et al. propose a more ambitious type of competition, which they call an "end-to-end causal competition," inspired by competitions in the field of bioinformatics. The idea is to prepare a true experimental study, while blinding the competitors (and possibly organizers as well) to its results. The competitors will have access to previously available observational data, and will have to estimate the result of the experiment. Access to domain experts will have to be made available to all competitors.

We believe this is an admirable suggestion for enlarging the scope of competitions such as the one we conducted. We will be glad to see members of the community take on this challenge.

7. CARNEGIE DISCUSSION

We appreciate Carnegie's comments about the positive externalities of our crowdsourced evaluation approached, namely reproducibility and generalizability. We hope others can capitalize on these strengths as she has.

Carnegie's discussion describes her valuable investigation of the impact of varying several features of the BART algorithm: TMLE, use of multiple chains, cross-validation to choose hyperparameters, including the propensity score as a covariate, and symmetric intervals. We hope that these preliminary results lead to a full-scale study. Her discussion shows that the posterior distribution of causal estimates depends critically on tuning parameters (the number of chains) as well as the inclusion of a propensity score estimate and that the addition of TMLE is less helpful than originally thought. As a case in point, with an optimized BART approach (multiple chains, propensity score included as a covariate, symmetric intervals used), the costs and benefits of TMLE correction become more clear. In particular, in this case adding TMLE yields a slight increase in coverage (91.9% to 92.7%) however at the cost of a substantial increase in RMSE and coverage interval length (0.016 to 0.022 and 0.04 to 0.07, respectively). While there are still benefits from merging these approaches, her discussion shows it may be possible to develop a "good enough" BART algorithm with minimal run time.

At a broader level, Carnegie's work represents an important case study highlighting the ways in which the competition data can be used yield new insights. In the development of easy-to-use black-box algorithms for causal inference, there remains room for improvement. The data from the competition provide a convenient test bed for pioneering and fine-tuning new developments.

8. FINAL THOUGHTS

Overall we feel that we have learned a tremendous amount by (1) running the competition, (2) analyzing the results, (3) exploring the efficacy of combining features of the highest-performing methods, and (4) reenvisioning the academic culture around the science of methodological development. We are confident that there is still more to be learned. We describe some key take-away messages in this last section.

8.1 Learning What *Not* to do

While our paper focused on the winning methods, given the repeated theme among the discussants that our competition focused on ideal circumstances, it is worthwhile to highlight that some very popular methods performed poorly overall, even in the DIY competition where researchers could take the lead on implementation.

8.2 Anticipating Future Competitions

Most discussants wished that the competition were broader. However it is worth noting that even the "small amount" of the landscape that we covered reflected a huge amount of work in order to make sure that the simulations satisfied all stated assumptions and were sufficiently challenging without being unsolvable. And this ignores the time spent managing the submissions, evaluating the performance, and analyzing those results. In other words if we want to encourage others to take on this responsibility in the future we should keep the bar of which each competition might accomplish at a reasonable level.

While the discussants often provided constructive criticism, only a few elaborated on the details of how their method was implemented or how they approached the problem. Future competitions might address this more directly for instance by creating opportunities for participants to include a deeper "reveal" about their choices (e.g., some notebook that maintains a log of choices/procedures, shared *a posteriori*).

There have been two similar competitions, both associated with the Atlantic Causal Inference Conference. BART was a top performer in both of these but the results have yet to published for either (rather results were announced at each conference). We are sympathetic given our understanding of the time involved—we hope funders will be willing to fund such work in the future to help support researchers who take on this task.

8.3 What More Could We Learn?

In running the competition, we were reminded of *how hard* it is to create simulations that are realistic but challenging and that satisfy certain constraints but not others. This is not a simple exercise and took an enormous number of hours to get right. We applaud all authors of creative simulation papers who have come before us as they inspired and informed our work.

Analyzing results is in some sense a bread-and-butter task for scholars with our proclivities and skill

sets. Since we designed the "experiment" we didn't even have to worry about selection bias! Unfortunately, just as with many data analyses, our results weren't as straightforward as we might have preferred. Maybe this means there exist a handful of methods that all perform well even in challenging circumstances. Maybe this is simply a reflection that the circumstances were too ideal. This merits further investigation.

8.4 New Methods Have Been Created

A clear win from this exercise is the creation of a host of new approaches that combine important features of some of the originally submitted high-performing methods, and some new features. For example, BART is now available in the SuperLearner library. In addition, augmented versions of BART (with multiple chains, new interval calculations, the estimated propensity score added as a covariate, and with TMLE adjustments) have been created and tested. These variations are currently available in the new bartCause package in CRAN.

One primary motivation behind the creation of the competition was dissatisfaction with the prevailing culture around how we evaluate the efficacy of methodological approaches to causal inference. We believe that team-based approached with crowd-sourced entries allowed us as researchers to take a more dispassionate stance towards performance, which in turn facilitated a more creative approach to methodological development

As a final point, we re-emphasize that we have only scratched the surface of the insights that might be gleaned from these datasets (Dorie, 2017). The datasets can be accessed on Dorie's github repository https://github.com/vdorie/aciccomp and we hope that more researchers choose to capitalize on this resource in their research.

ACKNOWLEDGMENTS

This research was partially supported by Institute of Education Sciences Grants R305D110037 and R305B120017.

REFERENCES

DORIE, V. (2017). aciccomp2016: Atlantic Causal Inference Conference Competition 2016 Simulation. R package version 0.1-0.

LAZER, D., KENNEDY, R., KING, G. and VESPIGNANI, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science* **343** 1203–1205.