Statistical Science
2019, Vol. 34, No. 1, 69–71
https://doi.org/10.1214/18-STS684
Main article: https://doi.org/10.1214/18-STS667

Comment: Spherical Cows in a Vacuum: Data Analysis Competitions for Causal Inference

Miguel A. Hernán

Abstract. A recent data analysis competition compared the performance of several methods for causal inference from observational data. However, sound causal inference requires not only adequate data analysis techniques but also subject-matter expertise about the causal structure of the problem under study. Therefore, until a methodology is developed to combine data analysis and subject-matter knowledge, causal inference competitions may only provide advice to practitioners under ideal conditions.

Key words and phrases: Causal inference, data analysis competitions.

We must be grateful to the organizers of the "2016 Atlantic Causal Inference Competition" for conducting the first large-scale data analysis competition for causal inference from observational data (Dorie et al., 2019). Modeled upon the machine learning competitions that are popular among data scientists, the 2016 competition challenged participants to estimate the causal effect of a treatment on an outcome using simulated datasets with many of the complications often found in real data.

Specifically, the participants in the 2016 competition were asked to apply their favorite method to a hypothetical subset of twin pregnancies from the Collaborative Perinatal Project (Niswander and Gordon, 1972). The goal was to estimate the average causal effect—defined as the effect in the treated—of low birth weight on child's IQ. The participants were provided with datasets generated under 77 different scenarios. The data generating processes for these scenarios differed with respect to degree of nonlinearity, overlap, percent treated, alignment between treatment and outcome models, treatment effect heterogeneity, and magnitude of treatment effect. Each dataset included 4802 observations, a binary treatment variable, a continuous outcome variable, and 58 covariates (not all of them confounders).

The 2016 competition had two flavors, labeled as "do-it-yourself" for participants who implemented their methods themselves and "black box" for participants who submitted a version of their method for implementation by the organizers. Participants proposed analyses based on stratification, matching (with and without propensity scores) and weighting. The submitted proposals included various modeling extensions and variable selection via machine learning algorithms. Several methods showed to be viable options that yielded relatively unbiased results. Methods that included flexible models for outcomes (regardless of whether they also modeled treatment) were the best performers, while methods that relied only on treatment prediction were at a disadvantage because adjustment for nonconfounders resulted in imprecise estimates.

While the "2016 Atlantic Causal Inference Competition" was an impressive exercise that explored the relative strengths and weaknesses of methods across a broad range of data generating processes, no competition can address all key challenges that investigators encounter when attempting causal inferences using observational data. For example, all simulated datasets had a dichotomous treatment (even though birth weight is actually a continuous variable), a continuous outcome, i.i.d. data, equal sample size, equal number of covariates and no measurement error. As Dorie et al.

Miguel A. Hernán is Professor, Departments of Epidemiology and Biostatistics, Harvard T.H. Chan School of Public Health, and Harvard-MIT Division of Health Sciences and Technology, Boston, Massachusetts 02115, USA (e-mail: miguel_hernan@post.harvard.edu).

(2019) explain, these restrictions were necessary to limit the number of scenarios provided to participants.

Like the organizers of the 2016 competition, I hope that others will be inspired to launch competitions that go beyond the restrictions (sensibly) imposed in 2016. In particular, I look forward to future competitions that include failure time outcomes and time-varying treatments and covariates. Note that the 2016 competition reflected the simplified aim of a large part of the causal inference literature: estimating the average causal effect of a time-fixed treatment. However, this emphasis on time-fixed treatments is at odds with the widespread presence of time-varying treatments and confounders in the health and social sciences. Future competitions may challenge researchers to use observational data with time-varying treatments and treatment-confounder feedback for the comparison of treatment strategies that are sustained over time (Robins, 1986). In these settings, it is likely that g-methods—g-formula, inverse probability weighting, g-estimation, and their doubly robust versions (e.g., TMLE)—will outperform many of the methods proposed in the 2016 competition, which cannot appropriately handle time-varying treatments with treatment-confounder feedback.

The experience accumulated during the "2016 Atlantic Causal Inference Competition" is a solid foundation for the extension of future competitions to more technically complex scenarios. Yet, regardless of their degree of technical sophistication, data analysis competitions will necessarily result in limited recommendations for applied researchers. The reason is that there is a fundamental mismatch between causal inference and data analysis competitions: causal inference from observational data requires not only adequate methods for data analysis but also sound subject-matter knowledge about the causal structure of the problem at hand (Hernán, Hsu and Healy, 2019). Let us see three examples of the reliance of causal inference on subject-matter knowledge.

First, adjustment for some variables may introduce systematic bias. These variables, which can often be be identified via subject-matter knowledge, may be colliders, mediators of the effect of treatment on the outcome, or instruments (which can amplify bias due to unmeasured confounders when, unlike in the 2016 competition, there are unmeasured confounders) (Greenland, Pearl and Robins, 1999; Pearl, 2011). However, none of the 2016 competition scenarios included variables that, if adjusted for, would induce bias in large samples. There is a good reason for the omission of biasing variables in a data analysis competition: in general, biasing covariates are statistically indistinguishable from debiasing covariates (confounders). Because no data analysis method can ever guarantee that biasing variables will be excluded from the adjustment set, little would be learned by including them in a data analysis competition.

Second, all scenarios in the 2016 competition included enough information to identify the causal effect of interest. That is, all scenarios were simulated under exchangeability of the treated and the untreated and positivity of treatment (also referred to as ignorability and overlap). Again, the organizers had little choice here. Suppose they had simulated scenarios in which important confounders were omitted from the dataset. In the absence of data on those confounders, none of the data analysis methods would have been able to correctly identify the causal effect. For example, suppose that birth weight is associated with IQ not because it causally affects IQ but because it is a marker for harmful intrauterine events that affect IQ. If a variable for those harmful intrauterine events is not in the dataset, all methods in the competition will fail to report the (truly) null effect of birth weight. Therefore, little would have been learned by including scenarios with unmeasured confounding.

Third, all scenarios in the 2016 competition assumed that the selection of individuals into the dataset was not in itself a source of bias. Suppose the organizers had simulated scenarios in which, unknown to the competition participants, children with missing IQ measurements were excluded from the dataset and in which the probability of having a missing IQ measurement depended on both maternal IQ and low birth weight. Then selection into the data (non-missing IQ) would be a conditioned on collider and selection bias would be expected (Hernán, Hernández-Díaz and Robins, 2004). However, no data analysis methods could have detected this bias.

An implication of the above is that starting point of all future data analysis competitions for causal inference may need to be somewhat unrealistic: datasets that do not include any biasing variables, that include all confounders, and that are not already conditional on colliders that would introduce selection bias.

Data analysis competitions are better suited for prediction than for causal inference. A prediction exercise may pay only a small price for neglecting subject-matter causal expertise once the question has been articulated and high-quality data have been obtained in

the population of interest. On the other hand, a realistic causal inference exercise cannot ignore subject-matter expertise because confounder identification and selection is riskier when not guided by expert knowledge.

The organizers of the 2016 competition were well aware of these limitations. As Dorie et al. (2019) note, the identification of unmeasured confounders "would require a great deal of subject matter expertise", which raises the question of how to fairly compare data analysis methods when their performance depends on the subject-matter expertise of the team that implemented them. Indeed, in the real world outside of data analysis competitions, research groups with subject-matter expertise are generally better positioned to make valid causal inferences. Dorie et al. (2019) state that "future competition organizers might consider ways to violate the key assumptions of ignorability and overlap for the inferential group." That will be a formidable task for causal inference competitions, which will have to find a way to satisfactorily combine data analysis and subject-matter expertise.

In summary, reducing a causal inference competition to a data analysis exercise may be necessary to learn about the statistical performance of various methodological approaches under artificial conditions. However, the overemphasis on adjustment and modeling techniques, at the expense of subject-matter expertise, results in a competition that is somewhat detached from the practice of causal inference in the health and social sciences. Researchers who look at causal inference competitions for practical advice are like the proverbial farmer who asked the local university for assistance to increase the milk production of his cows. A sophisticated theoretician wrote back "I have the solution, but it works only in the case of a spherical cow in a vacuum."

ACKNOWLEDGEMENTS

This work wassupported in part by NIH Grant R37 AI102634. The last sentence was retrieved on November 6, 2018 from https://en.wikipedia.org/wiki/Spherical_cow (Spherical cow, 2018).

REFERENCES

GREENLAND, S., PEARL, J. and ROBINS, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology* **10** 37–48.

HERNÁN, M. A., HERNÁNDEZ-DÍAZ, S. and ROBINS, J. M. (2004). A structural approach to selection bias. *Epidemiology* **15** 615–625.

HERNÁN, M. A., HSU, J. and HEALY, B. (2019). Data science is science's second chance to get causal inference right. A classification of data science tasks. *Chance* **32** 42–49.

DORIE et al. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statist. Sci.* **34** 43–68.

NISWANDER, K. R. and GORDON, M. (1972). *The Collaborative Perinatal Study of the National Institute of Neurological Diseases and Stroke*: *The Women and Their Pregnancies*. W.B. Saunders, Philadelphia, PA.

PEARL, J. (2011). Understanding bias amplification. *Am. J. Epidemiol.* **174** 1223–1227.

ROBINS, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods-application to control of the healthy worker survivor effect. *Math. Model.* **7** 1393–1512. (Errata appeared in *Comput. Math. Appl.* **14**(1987), 917–921).

SPHERICAL COW. In Wikipedia. Retrieved November 6, 2018, from. https://en.wikipedia.org/wiki/Spherical_cow.