Statistical Science
2019, Vol. 34, No. 1, 86–89
https://doi.org/10.1214/18-STS679
Main article: https://doi.org/10.1214/18-STS667

Comment: Causal Inference Competitions: Where Should We Aim?

Ehud Karavani, Tal El-Hay, Yishai Shimoni and Chen Yanover

Abstract. Data competitions proved to be highly beneficial to the field of machine learning, and thus expected to provide similar advantages in the field of causal inference. As participants in the 2016 and 2017 Atlantic Causal Inference Conference (ACIC) data competitions and co-organizers of the 2018 competition, we discuss the strengths of simulation-based competitions and suggest potential extensions to address their limitations. These suggested augmentations aim at making the data generating processes more realistic and gradually increase in complexity, allowing thorough investigations of algorithms' performance. We further outline a community-wide competition framework to evaluate an end-to-end causal inference pipeline, beginning with a causal question and a database, and ending with causal estimates.

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

Scientific challenges, in the form of competitions, have gained popularity in recent years. Nowhere is this more evident than in the field of machine learning. What started as one-time challenges (e.g., Netflix Prize, 2006; Bennett, Lanning et al., 2007), evolved into a major part of many annual conferences (such as ImageNet Large Scale Visual Recognition Challenge, 2010-present; Russakovsky et al., 2015) and further developed into vibrant ongoing platforms (e.g., Kaggle and Synapse). These competitions act as remarkable growth engines for the field of machine learning by boosting tool development and pushing forward new ideas and algorithms. This is accomplished, in part, by engaging the community and attracting researchers from various disciplines, who further diversify both the problems and solutions.

Ehud Karavani is researcher, Healthcare Informatics Department, IBM Research—Haifa, Israel (e-mail: ehudk@ibm.com). Tal El-Hay is researcher, Healthcare Informatics Department, IBM Research—Haifa, Israel (e-mail: talelh@il.ibm.com). Yishai Shimoni is researcher, Healthcare Informatics Department, IBM Research—Haifa, Israel (e-mail: yishais@il.ibm.com). Chen Yanover is researcher, Healthcare Informatics Department, IBM Research—Haifa, Israel (e-mail: cheny@il.ibm.com).

Therefore, adopting crowdsourcing concepts and importing data challenges into the causal inference domain can be highly beneficial. They may establish agreed-upon benchmarks, promote fair comparisons between the abundance of algorithms using standardized measures, drive the development of ever-improving methods, and draw new researchers to the domain.

However, constructing causal inference data challenges is in itself challenging. This is due primarily to "the fundamental problem of causal inference" (Holland, 1986): it is impossible to observe the results of two different courses of action (or "interventions") on the same entity and, consequently, it is impossible to observe their comparative effect. And when such "ground truth" is unavailable, it is unclear how to evaluate the performance of submitted results.

One way to combat this problem is using synthetic data, for which we can explicitly simulate outcomes for all potential interventions, and so directly compute the comparative effects and test counterfactual predictions. These provide a reasonable framework to evaluate an algorithm's performance and are the basis of current challenges. Although simulation-based competitions don't cover all aspects of the causal inference pipeline in practice, they are a highly important first

step in the adoption of competitions by the causal inference community. In this note, we elaborate on some missing ingredients of a simulation-based framework and propose a potential avenue to address them.

SIMULATION-BASED COMPETITIONS: WHAT'S MISSING?

Despite their tremendous potential, simulation-based competitions are still missing three main ingredients worth addressing.

Lacking evaluation of the entire causal workflow. Competitions, in their current form, only compare algorithms in a sand-box manner; that is, the cohort, treatment assignment, covariates and outcomes are precomputed by the competition organizers and given to participants. Meanwhile, the day-to-day focus of a causal inference researcher involves much more than simply executing a chosen algorithm. Given a research question, a preliminary nontrivial (Jones, Molitor and Reif, 2018, Silberzahn et al., 2017) and essential task involves specifying the target trial protocol (Hernán and Robins, 2016) and extracting its components from an available observational database. Next, the researcher must characterize the underlying causal structure (and extract the related covariates), based on either domain expert knowledge or data-driven methodologies. Only then do causal inference algorithms come into play. While testing just the latter part has its own merit, we argue that a competition providing a causal question and a database can be more comprehensive and truly research-oriented than the current approach.

Disregarding the underlying causal structure. Current competitions also deprive the extracted covariates from having a meaning. Specifically, the covariates used for simulating the treatment assignment and outcomes are either masked (as in the ACIC 2016 and 2017 competitions) to avoid revealing their true identity (as it is disregarded anyway), or the causal structure itself is randomized (ACIC 2018 competition). This leaves the participants with two possible plans of operation. They either rely on automated causal discovery methods to identify a minimal set of confounders (as far as we know, such an approach was not applied in the competitions to date). Or they condition on *all* covariates, a reasonable practice under the *in-silico* assumptions of past competitions, but nonetheless unfavorable for real world scenarios as it may bias estimations and increase their variance.

Adhering to strict assumptions. Simulations are only as good as their underlying assumptions, and current competitions rely on multiple assumptions that may not be correct in real life. A comprehensive list of assumptions worth relaxing and potential extensions to future competitions can be found in Dorie et al. (2017), Section 4.4. These include many widely used assumptions such as strong ignorability, and data samples being independent and identically distributed (IID) as well as simulation attributes such as nonbinary treatments. Additional augmentations that we find very important are the inclusion of simulated covariates to act as mediators, and censored outcome variables (as in missing values) arising from an informative mechanism; these are found in a plethora of real-world examples. Elaborate treatment assignment mechanisms, such as "targeted selection" introduced in the 2017 competition (Hahn, Dorie and Murray, 2018), multiple treatments (both overlapping and mutually exclusive), and time-varying treatments, are also worth examining. It is also essential to incorporate effect on multiple outcomes, ideally combining different types, scales, and units, and under different kinds of potential biases. Various noise models should also be considered, and introducing these may call for additional evaluation metrics, as suggested by Shimoni et al. (2018).

Extended Simulation-Based Competitions

Despite the above disadvantages, simulation-based competitions are still a valuable tool for thoroughly investigating causal algorithms. The control over each parameter of the data generating process (DGP) and the vast number of datasets it is able to produce, enable researchers to gain a better understanding of the model's uncertainty, weak spots, and other trade-offs.

However, a full understanding can only be achieved by gradually relaxing the simulation assumptions. Each competition should test specific issues (such as IID violations or performance on categorized responses), carefully controlling its DGPs' parameters to vary its effect (e.g., gradually increasing cohort size).

We refer to these as a kind of dose-response tactic, where the *dose* is the tweaking of different DGP parameters and the *response* is any evaluation metric of the algorithm's performance (accuracy and precision, but also meta-metrics such as running time). This tactic could contribute a great deal to our causal understanding of the practical scenarios where certain algorithms are preferable to others. As an example, censored outcome and data scalability (in terms of cohort size) were addressed in the 2018 competition. Unfortunately, no meaningful conclusions could be drawn, due to the low number of participants and missing textual description

of the submitted methods including user-provided run times.

We also advocate a reconsideration of the scoring schemes to better capture the performance of each submitted algorithm. Specifically, scoring each DGP (or a set of conceptually similar DGPs) separately may highlight algorithms that are superior for some specifications of causal structures, a signal easily lost when averaging over multiple different structures. Previous competitions did not pin-point a single winner. They identified a group of top performers or downplayed the competitive aspect of the challenge altogether. If, as a community, we do want to recognize the winner(s), it is advisable to announce a ranking function ahead of time, and to reasonably (even if arbitrarily) combine the applied metrics (potentially of different units, like bias and coverage). An example for such an approach could rank per-DGP performance and later aggregate over all DGPs.

Lastly, online platforms for automating the process of submission and scoring are highly beneficial. They hold a few key advantages, which led the 2018 competition to adopt them. For organizers, they lighten the burden of manually scoring submissions, thereby reducing fault-prone human involvement. For participants, they increase trust by being both transparent and server-operated, and can provide feedback in real time. The cost of using such platforms is the higher barrier-to-entry for participants, who must learn to navigate that platform. This may deter some for starters. However, as the platforms improve and researchers become more aware of open community research, we believe the transition will occur.

END-TO-END CAUSAL COMPETITION: IS IT FEASIBLE?

Although the augmentations just proposed can partially address the issue of DGPs adhering to strict assumptions, simulation-based competitions still serve as a tool to compare algorithms only, while the underlying causal structures are typically disregarded. We describe below a challenge framework that evaluates the entire causal inference pipeline in a real-world setting, as opposed to only the effect estimation under certain, not necessarily realistic, assumptions.

The proposed framework is inspired by experiments like the Critical Assessment of protein Structure Prediction (CASP) (Moult et al., 2018) and Critical Assessment of PRediction of Interactions (CAPRI) (Lensink, Velankar and Wodak, 2017). There, information about soon-to-be-solved three dimensional protein structures is collected from the experimental community. Amino acid sequences of these proteins are provided to participants who, in turn, model (and submit) their structure. Finally, independent assessors blindly analyze, score, and interpret the submitted structures using established measures against the newly solved structure.

We envision a similar approach being taken to support end-to-end causal challenges. Specifically, a community-nominated entity will collect information on soon-to-be-published comparative effectiveness "experimental" trials, that can be reproduced using nonexperimental, observational data. Such trials include: constructed observational studies, combining experimental (typically, on treated entities) and observational (untreated ones) data (Hill, Reiter and Zanutto, 2004), studies conducting randomized and nonrandomized trials in parallel (Shadish, Clark and Steiner, 2008), and experimental trials that can be emulated using available observational data (Danaei et al., 2018).

On a designated date, for a limited time window, participants will be granted access to two sources: trial information to define the target trial's protocol (Hernán and Robins, 2016) and observational data to extract the protocol's components, namely the study cohort, treatment assignment, and outcomes. Domain expert knowledge, as well as data-driven methodologies (Malinsky and Danks, 2018), may then be used to infer a causal structure and identify a (minimal) set of confounding factors. Finally, participants will estimate the causal contrast of interest. Depending on the experimental study design, these may include the effect of the studied treatment(s) on several outcomes and in different sub-populations, to avoid concentrating an entire competition on evaluation of a single number. Estimation accuracy will be blindly evaluated by independent assessors against the results of the controlled trial.

Since an end-to-end causal inference pipeline contains many different steps, it might be challenging to draw strong conclusions based only on the accuracy of the estimated contrasts. To facilitate better comparison of the evaluated pipelines, teams will be encouraged to submit modular and well-documented entries. Each well-defined module will correspond to a specific step of the process (e.g., covariate extraction, causal graph construction, and causal algorithm), or an intermediate result (for example, expert-identified confounders). These submitted components could be later mix-and-matched across teams (as was implemented in the post-competition methods by the organizers of the

2016 competition) to obtain more conclusive insights on the performance of each approach.

A concerted community effort, potentially spanning multiple research domains, can highlight the strengths and weaknesses of current approaches for causal effect estimation from observational studies, in a real-world setting. Importantly, as demonstrated by the CASP and CAPRI experiments, it may lead to significant advances in causal estimation and discovery, and expose this line of research to wider more diverse audiences.

CONCLUSIONS

Simulation-based competitions are an excellent tool for the large scale assessment of algorithms. By having full control over the data generating processes, it is possible to examine algorithm performance in total isolation under minute changes and over multiple instances. On the down side, such competitions typically adhere to unrealistic assumptions, lack real-world meaning, and are missing a big portion of the scientific process.

In contrast, end-to-end competitions can uncover the differences between the analysis power of observational data and actual randomized controlled trials (RCTs), but only for a handful of instances at a time. Practically speaking, they are limited by the amount of RCTs occurring at a given time and can be much more demanding to work on, thereby increasing the burden on the participants.

We would like to end with a call for organizers and participants to take part. Only through participation can we establish a diverse pool of tools to check, and to gain enough statistical power to extract valid insights from their performance.

ACKNOWLEDGMENTS

We thank Prof. Ashley I. Naimi, University of Pittsburgh, for leading the ACIC 2018 data challenge. We are grateful to Omer Weissbrod for his insights and to the reviewer for the constructive inputs. Lastly, we wish to thank the researchers from the Machine Learning for Healthcare and Life Sciences team, IBM Research—Haifa for fruitful discussions.

REFERENCES

BENNETT, J., LANNING, S. et al. (2007). The Netflix prize. In *Proceedings of KDD Cup and Workshop* **2007** 35. New York, NY.

DANAEI, G., RODRÍGUEZ, L. A. G., CANTERO, O. F., LOGAN, R. W. and HERNÁN, M. A. (2018). Electronic medical records can be used to emulate target trials of sustained treatment strategies. *J. Clin. Epidemiol.* **96** 12–22.

DORIE, V., HILL, J., SHALIT, U., SCOTT, M. and CERVONE, D. (2017). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. ArXiv Preprint. Available at arXiv:707.02641.

HAHN, P. R., DORIE, V. and MURRAY, J. S. (2018). Atlantic Causal Inference Conference (ACIC) Data Analysis Challenge 2017.

HERNÁN, M. A. and ROBINS, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *Am. J. Epidemiol.* **183** 758–764.

HILL, J. L., REITER, J. P. and ZANUTTO, E. L. (2004). A comparison of experimental and observational data analyses. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. *Wiley Ser. Probab. Stat.* 49–60. Wiley, Chichester. MR2134801

HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–970. MR0867618

JONES, D., MOLITOR, D. and REIF, J. (2018). What do workplace wellness programs do? Evidence from the Illinois workplace wellness study. Technical report, National Bureau of Economic Research, Cambridge, MA.

LENSINK, M. F., VELANKAR, S. and WODAK, S. J. (2017). Modeling protein-protein and protein-peptide complexes: CAPRI 6th edition. *Proteins* **85** 359–377.

MALINSKY, D. and DANKS, D. (2018). Causal discovery algorithms: A practical guide **13** e12470.

MOULT, J., FIDELIS, K., KRYSHTAFOVYCH, A., SCHWEDE, T. and TRAMONTANO, A. (2018). Critical assessment of methods of protein structure prediction (CASP)—Round XII **86 Suppl. 1** 7–15.

RUSSAKOVSKY, O., DENG, J., SU, H., KRAUSE, J., SATHEESH, S., MA, S., HUANG, Z., KARPATHY, A., KHOSLA, A. et al. (2015). ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115** 211–252.

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

SHIMONI, Y., YANOVER, C., KARAVANI, E. and GOLDSCHMNIDT, Y. (2018). Benchmarking framework for performance-evaluation of causal inference analysis. ArXiv Preprint. Available at arXiv:1802.05046.

SILBERZAHN, R., UHLMANN, E., MARTIN, D., ANSELMI, P., AUST, F., AWTREY, E., BAHNIK, Š., BAI, F., BANNARD, C. et al. (2017). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. PsyArXiv.