Biostatistics (2019) **0**, 0, *pp.* 1–3
doi:10.1093/biostatistics/kxz045

Machine learning for causal inference in *Biostatistics*

SHERRI ROSE*

Department of Health Care Policy, Harvard Medical School, 180 Longwood Ave, Boston, MA, 02115, USA

rose@hcp.med.harvard.edu

and

DIMITRIS RIZOPOULOS

Department of Biostatistics, Erasmus University Medical Center, PO Box 2040, 3000 CA Rotterdam, the Netherlands

General inference problems and quantifying uncertainty have long been the cornerstone of statistical science. While machine learning advances have permeated many disciplines, inference for these procedures, and in particular, causal inference, has not been widespread. However, this is rapidly changing. As different scientific fields begin to converge on machine learning for causal inference, we thought now would be an excellent time to have a public discussion. In our roles as editors of *Biostatistics*, we decided to organize a series of commentaries on the topic from scholars with expertise in statistics, computer science, epidemiology, health economics, policy, and law. We intentionally invited leaders who are early or mid-career scholars and considered multiple dimensions of intersectional diversity to increase the range of voices given a platform. Machine learning *specifically for causal inference* is a smaller area, thus this involved reading conference programs, arXiv papers, and department websites along with sending invitations in waves in an attempt to achieve a balance of perspectives. Not everyone said yes, which is not unexpected given we were deliberately reaching outside our professional networks and across disciplines. We share these experiences for the potential benefit of other organizers and to argue that this time investment is necessary. The collection we curated contains five pieces that we briefly introduce here.

The first commentary discusses the centrality of understanding structural racism when implementing machine learning (Robinson *and others*, 2019). Structural racism is pervasive in health applications, and the authors expertly present their thesis through the use of causal graphs. Causal modeling forces researchers to think critically about how their data were generated and also allows an enriched causal interpretation of their statistical target parameter. Assessing and eliminating algorithmic bias is a growing area of research, but most efforts do not focus on health and biomedicine. We encourage scholars to engage in these issues and give them consideration in each project. This should be required when creating tools meant to be deployed or making policy recommendations.

In Subbaswamy and Saria (2019), the authors address the key topic of generalizability. A lack of generalizability for a given algorithm can be due to many factors, including shifts in conditions between the training scenario and the system where it was applied. A serious treatment of generalizability is needed

*To whom correspondence to be addressed.

for questions in causal inference (and also prediction, among others). The authors leverage causal graphs to understand the underlying data generating process, assess susceptibility to shifts, and summarize some algorithmic approaches. Particularly as we see machine learning begin to proliferate the clinical and public health literatures, it is important that we set a standard to address generalizability concerns in all work making claims that it is ready for broad practical use.

The commentary from Diaz (2019) is the first in this series to deal primarily with estimation. While the causal modeling step prior to estimation is given clear emphasis, the unification of machine learning and inference is the dominant subject. The author focuses on procedures that rely on semiparametric theory in order to build machine learning-based effect estimators that can, when combined with additional causal assumptions, produce causal estimates. Specifically, the properties of targeted minimum loss-based estimation (TMLE) and double/debiased machine learning (DML) are enumerated. TMLE and DML represent successes from the literature where data-adaptive procedures can be used to estimate relevant nuisance parameters while still obtaining valid statistical inference.

The promise of precision medicine and identifying individualized treatments is of growing interest in clinical settings. Shalit (2019) explores the estimation of individual-level treatment rules. The ability to estimate these effects and with what data are much debated. The author dives into this debate, presenting statistical and causal challenges as well as illuminating misunderstandings surrounding concerns of p-hacking. Sample sizes in randomized controlled trials are a major limitation, but observational data may yield insights if proper control of confounding and other matters can be addressed. Fusion of randomized and observational data is another potential avenue.

We close the series with a commentary on regulatory and policy issues (Stern and Price, 2019). The integration of machine learning in health care has taken many forms, including decision support and testing, and we now have devices with machine learning-based software. The FDA regulates so-called "software as a medical device," but there are distinct complications that arise in understanding their safety and effectiveness. The challenge of regulating algorithms is enormous and ongoing, especially as collecting digital biomarkers increases. The authors also nicely tie together themes from the previous four commentaries (generalizability, algorithmic bias, machine learning-based effect estimation, and individualized treatments), demonstrating how they all appear when examining regulation and policy in the machine learning space.

We by no means claim to have included all the important facets of or viewpoints on machine learning for causal inference. There are many more, but we hope this series of commentaries helps to spur discussions across disciplines. A key theme we stress is the thoughtful consideration required when pursuing this line of work. Machine learning for causal inference in health and biomedicine should not be treated as a fad to join or mere proving ground for new algorithms given the stakes involved. To quote Nick Jewell, we need to remember that "behind every data point there is a human story, there is a family, and there is suffering" (Jewell, 2003). Genuinely engaging in the complexities of the applied real-world problems, theoretical underpinnings, and potential social impacts, particularly for marginalized groups, is crucial.

Acknowledgments

Conflict of Interest: None declared.

References

Diaz, I. (2019). Machine learning in estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics*. doi:10.1093/biostatistics/kxz042.

Jewell, N. (2003). *Statistics for Epidemiology*. Boca Raton, FL: Chapman and Hall/CRC.

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.

SHALIT, U. (2019). Can we learn individual-level treatment policies from clinical data? *Biostatistics.* doi:10.1093/biostatistics/kxz043.

STERN, A. D. AND PRICE, W. N. (2019). Regulatory oversight causal inference, and safe and effective health care machine learning. *Biostatistics*. doi:10.1093/biostatistics/kxz044.

SUBBASWAMY, A. AND SARIA, S. (2019). From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics*. doi:10.1093/biostatistics/kxz041.

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