American Journal of Epidemiology

Vol. 191, No. 10
https://doi.org/10.1093/aje/kwac108
Advance Access publication:
June 28, 2022

Commentary

The Future of Causal Inference

Nandita Mitra*, **Jason Roy, and Dylan Small**

* Correspondence to Dr. Nandita Mitra, Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, 423 Guardian Drive, Philadelphia, PA (e-mail: nanditam@upenn.edu).

Initially submitted May 1, 2022; accepted for publication June 17, 2022.

The past several decades have seen exponential growth in causal inference approaches and their applications. In this commentary, we provide our top-10 list of emerging and exciting areas of research in causal inference. These include methods for high-dimensional data and precision medicine, causal machine learning, causal discovery, and others. These methods are not meant to be an exhaustive list; instead, we hope that this list will serve as a springboard for stimulating the development of new research.

algorithms; causal discovery; causal machine learning; distributed learning; high-dimensional data; interference; transportability

Abbreviation: DAG, directed acyclic graph.

Editor's note: *The opinions expressed in this article are those of the authors and do not necessarily reflect the views of the* American Journal of Epidemiology.

Over the past few decades there have been major achievements in the development of causal inference theory and methods and in a range of applications. Foundational advancements in modern causal inference have come from diverse fields, including epidemiology, biostatistics, statistics, computer science, and economics. Seminal work by James Heckman, Judea Pearl, James Robins, Paul Rosenbaum, and Donald Rubin (among others) led to groundbreaking changes in how problems are approached and data are analyzed. For example, researchers who are well versed in causal inference ideas will typically take great care in defining the population of interest, specifying the target causal parameter(s), assessing identifying assumptions using subject matter knowledge (possibly with the help of directed acyclic graphs (DAGs)), designing the study to emulate a target trial, choosing efficient and robust estimators, and carrying out sensitivity analysis. In the past 40 years, novel approaches such as propensity scores, instrumental variables, mediation analysis, and methods for estimating optimal dynamic treatment regimens have been developed and are now commonly used by applied researchers to answer impactful questions.

As codirectors of the Center for Causal Inference (a joint partnership between the University of Pennsylvania and Rutgers University), we are naturally passionate about causal inference. It is heartening to see the exponential growth in causal inference methodology and the explosion of applications in medicine, education, sociology, and public policy. It has also been exciting to see the increase in interest in causal inference among our graduate students. When the 3 of us were in graduate school in the 1990s, there were very few formal courses in causal inference. Now, most doctoral programs in statistics, epidemiology, and biostatistics offer semester-long courses, and there is a nice cadre of excellent textbooks (1–3). There is also now the newly formed Society for Causal Inference (SCI), which will bring together causal researchers across different disciplines to foster research collaborations and enhance training opportunities.

We are delighted to have been invited by the Editor-in-Chief of the *Journal* to write a commentary on the future of causal inference. Based on our collective experiences and having polled our colleagues currently engaged in cutting-edge causal inference research, we provide here our top-10 list of new and exciting areas of research in causal inference. These are not listed in any specific order of importance

and are not meant to be a review of existing methods. Nor do we intend this list to be exhaustive, by any means. Instead, we hope that this list will serve as a springboard for stimulating the development of new research. The future of causal inference is bright. There is much to do and many, many bright new minds to do it.

TOP 10 FUTURE DIRECTIONS FOR CAUSAL INFERENCE RESEARCH

High-dimensional data

High-dimensional data are increasingly available for researchers in both randomized trials and observational studies. This opens possibilities to learn about a broader range of causal questions more reliably, but it also presents statistical challenges. High-dimensional data can come in many forms, including high-dimensional exposures, confounders, and mediators. In all these cases, statistical approaches could involve variable selection or other types of dimension reduction (along the lines of principal components analysis). Careful consideration needs to be made for the plausibility of the causal identifying assumptions as well as the interpretability. Further, in these high-dimensional data problems there might be tradeoffs between making strong assumptions on the dimension-reduction part of the problem or on the modeling part of the problem.

Arguably, the most common type of high-dimensional data problem in causal inference is high-dimensional confounding. Because valid causal inference often requires an "ignorability" type of an assumption, having access to a large amount of potential confounders is valuable. Variable selection problems here are more complex than in standard prediction modeling problems because, ideally, we will include variables that affect the treatment decision (propensity score) and the outcome. There is a small but growing literature on methods that take both types of relationships into account in variable selection (as discussed elsewhere (4–6)). Other types of dimension-reduction approaches run the risk of invalidating the ignorability assumption, although the bias-variance tradeoff might make this acceptable. Examples of high-dimensional exposures could include genetic variants (7), environmental exposures (8), or multidimensional treatments with some continuous components (9). It a distinct challenge to determine how one can reduce the dimension of the exposure while still having interpretable estimands that allow for possible exposure interaction effects. Recent motivating examples in the literature for high-dimensional mediation include genomics data and functional magnetic resonance imaging data as mediating variables. In addition to the challenges mentioned above in terms of defining causal effects and assessing the plausibility of identifying assumptions, high-dimensional mediation also has another challenge. Because the mediators can potentially be affected by the exposure(s), we might want to capture change in the mediators (pre/post exposure). Suppose the mediators are microbiome data or genomic data. How might we account for preexposure values and changes in these variables after treatment?

Precision medicine

Precision medicine is an idea that has received a lot of attention over the past decades, from excitement about the potential impact of the Human Genome Project to President Obama's Precision Medicine Initiative, through to the present day. It can be thought of as using available data to determine what treatment is best for an individual and delivering it at the right time. Although the idea of precision medicine was largely motivated by choosing medications or other medical interventions based on an individual's genomic data or other biomarkers, we can also think more generally, beyond just medicine, to tailored interventions and policies.

One of the key statistical questions for precision medicine is, "given a person's historical data (including prior treatments), what is the best course of action?" Breakthroughs in the statistical literature on estimating optimal dynamic (adaptive) treatment strategies occurred in the early 2000s (10, 11). Since then, we have seen many advances in statistical methodology, as well as some real-world applications of the methods. Some recent statistical developments have made it possible to allow frequent, irregular measurement times, and to get even more precise on when to monitor or deliver interventions. For example, taking into account the costs of obtaining data (such as the monetary cost of obtaining biomarker values or the cost of a person's time or interest in responding to survey questions), one could determine both the optimal monitoring plan and the optimal intervention given data obtained from that monitoring plan (12, 13). The opportunities for precision medicine are furthered using mobile devices (mHealth) in health care and research. Microrandomized trials are designed to take advantage of mobile technology to have real-time data collection, randomization, and delivery of interventions (14). While there have been some statistical advances for estimating dynamic treatment strategies from mHealth data (as in Luckett et al. (15)), we believe there will be a lot of development in the next decade. In addition, there is a particularly strong need here for team science (close collaboration with clinical researchers) and translation (software, training).

Causal machine learning

Causal inference approaches involve study design, defining causal estimands, identifying (causal) assumptions, and statistical modeling. In order to specify, for example, a propensity score model or an outcome model (or both) to make causal inference, we need to learn about observed data distributions or functions (such as mean functions). Machine learning methods allow analysts to avoid making strong parametric assumptions, potentially leading to a reduction in risk of bias due to model misspecification. Causal machine learning differs from standard machine learning in key ways. We are not trying to predict what will happen next given the way the world currently is. In causal machine learning, we are trying to predict what would happen if a particular aspect about the world changed. (For example, what if everyone in our population of interest followed a particular treatment plan?) Because of this, before machine learning methods are

implemented, careful thought needs to be given to the design of the study, what variables to include in which models, and so on. There is much to be done in this area of research.

Enriching randomized experiments with real-world data

Randomized experiments provide an unbiased way to estimate treatment effects by dividing units into treatment and control groups in an unbiased manner by "a flip of a coin". However, randomized experiments are expensive and difficult to conduct, and are often done only among subgroups of the population of interest. Observational studies, which rely on "real-world" data, have opposite advantages and disadvantages: They can be severely biased by confounding because people themselves rather than coins decide on who is in the treatment and control group, but they are less expensive and can be conducted considering the whole population of interest.

An important area of research is how to best combine evidence from observational studies and randomized trials in making causal inferences. This could be seen as a form of meta-analysis that combines evidence from different types of studies. Examples of how such combinations are useful in the context of studying the effect of hormone replacement therapy for postmenopausal women are given by Prentice et al. (16). An example of work on combining evidence from randomized experiments and observational studies is Kaizar (17), in which estimates were made of the treatment effect among the part of the population where the randomized experiment was conducted and then the observational study was used to estimate how different the part of the population among which the trial had been conducted was from the part of the population among which the trial had not been conducted. Another way in which real-world data can enrich randomized trials is by enabling improved trial design. An example is Shortreed et al. (18), which uses electronic health record data to make more realistic sample size calculations in a suicide prevention randomized trial.

Algorithmic fairness and social responsibility

Machine learning, deep learning, and artificial intelligence approaches are often used to develop clinical algorithms to aid with risk prediction and decision making for treating patients. Private industries, government agencies, and the criminal justice system also use these approaches to determine insurance rates and hiring practices, perform facial recognition, and predict recidivism when making sentencing recommendations (19). These tools, however, are only as good as the data that are used to develop them. Often, the data sources are biased and suffer from deep-rooted social and systemic inequities and injustices. In order to address these issues head-on, careful thought must be given to the underlying causal pathways that give rise to unfair practices, such as disparities in access to health care. Kusner and Loftus (20) lay out ways in which causal modeling may help assess the fairness or bias of such algorithms by thinking in terms of counterfactuals (e.g., would the prediction from the model change if we changed one feature of an individual, such as their race) and by conducting sensitivity analyses of the algorithms to assess whether they could be biased because of unknown or unmeasured factors. As machine learning approaches are being developed and improved to leverage massive amounts of data, in parallel, principled data collection and causal methods need to be developed to aid in the training and assessment of algorithms to mitigate hidden discrimination and unfair practices. As Kusner and Loftus recommend, epidemiologists and statisticians should work closely with ethicists, social scientists, clinicians, stakeholders (e.g., patients that the algorithm may affect), and others in an interdisciplinary fashion so that algorithms are based on rich and diverse data and incorporate critical features in the causal pathway.

Distributed learning

We have seen the term "distributed learning" used in several different ways, and all of it could end up being important for causal inference problems. For example, it often refers to distributing computational workload across many machines to help with scalability (21). Deep learning models might require an extremely large number of parameters and enormous amount of training data, and, if centralized, might not be feasible.

Distributed learning aims to distribute the workload while finding cohesive ways to integrate the information. As we discuss in the section on causal machine learning, machine learning methods in general will increasingly play a crucial role in causal inference. Another use of the term has to do with using multiple data sets that cannot be combined or merged. For example, several health-care systems might be willing to allow models to be fitted to their data and to share the output but not to share granular patient data. This kind of privacy-preserving approach will likely be increasingly common in the future. How best to go about causal inference when you can fit models to different data sets from potentially different populations, without having access to the raw data? Many of the other problems that we discuss in this article are also relevant here, such as transportability.

Causal discovery

Much of the causal inference literature begins with assumptions depicted by DAGs and proceeds with methods development given the DAG. Causal discovery, on the other hand, has to do with uncovering causal relationships and structures from observational data using computational and statistical methods. The DAG cannot be identified purely from the data (i.e., it cannot be identified purely from knowing the probabilistic relationships between the observed variables). For example, without any further assumptions, it cannot be known whether an association between X and Y arises from X causing Y, Y causing X, or neither causing the other and instead an unmeasured confounder generating the association. However, under certain assumptions, such as the causal Markov assumption or the causal faithfulness assumption, the DAG can be identified from the data (22). Under one of these assumptions, traditionally, score-based

methods have been used to discover DAGs (23). In this approach, the space of DAGs is searched for the optimal score using computational algorithms. We anticipate that causal discovery will be an area of substantial development in the coming years. New assumptions under which causal discovery can be undertaken that are motivated by specific applications (e.g., metabolomics, mental health disease pathophysiology) will be needed. Further, new methods may be needed for data gathered under complex conditions, such as having aggregate measurements of causal interactions at a finer scale or data that have a network structure (e.g., some people in the data are friends).

Interference and spillover

In the assessment of public policies, we are often concerned about how to account for the policy "spilling over" to nearby regions. For instance, coronavirus disease 2019 lockdown measures in one city may affect infection rates in nearby towns; residents of the suburbs of Philadelphia may work, go to restaurants, and attend concerts in Philadelphia, and hence, mandates in the city of Philadelphia will affect infection rates in the suburbs in ways that may not be measurable. On the other hand, residents of the suburbs may opt to dine locally rather than venture to Philadelphia. It is also possible that the suburbs may take Philadelphia's lead and adopt similar behaviors, such as social distancing and masking. Either way, the policy choices of these regions interfere with each other. Assessing the causal effect of spillover and defining relevant estimands of interest are emerging areas of research in causal inference.

When it comes to accounting for spillover, estimands of interest may include the average treatment effect on the treated in the presence of spillover and the average treatment effect on a neighboring control. For instance, there is interest in assessing the causal effect of excise taxes, such as sweetened beverage taxes, on consumption (24). However, cross-border shopping from a taxed region to a nontaxed neighboring region can mitigate the effect of the tax on the region of interest and also affect the neighboring region's volume of sales. Similar complexities arise in the assessment of the causal effects of neighborhood policing initiatives on crime (25). Here, spillover effects of one precinct's policies on neighboring precincts is critical in understanding the causal effects of such policies. Of course, spillover patterns or behaviors may be influenced by unmeasured confounders complicating the issue but potentially fueling future research. One can imagine an even more complex situation in which the causal effect of air pollution policies on health is of interest. Spillover effects, in this case, must also take into account spatiotemporal correlations (26). Causal identification and modeling of relevant spillover estimands in these situations are challenges that are yet to be tackled.

Finally, there is currently a vast amount of data being curated, leveraged, and wrangled from social media platforms such as Twitter (San Francisco, California) and Facebook (Menlo Park, California). For instance, there has been interest in studying the association between geographic variability in social media postings and diabetes rates (27). What if we wanted to assess the causal effects of social media postings on human health? Does the more someone tweets about their diabetes indicate more awareness of their own health as well as influencing others to lead healthier lifestyles? These are social networks that are subject to interference and network dependencies (28) but are also subject to unmeasured confounding (e.g., sociodemographic factors) and substitution (those who post a lot on Facebook may not post on Twitter or LinkedIn (Sunnyvale, California)).

Transportability

Decision makers are often interested in whether results of a study conducted on a specific population can be transported to another population of interest. This could be in the context of transporting, for instance, the causal average treatment effect from a randomized controlled trial to a different target population. Elegant causal theory and assumptions have been laid out in the setting of transporting causal conclusions from experimental settings to observational settings using a structural causal model framework (29). There may also be interest in transporting the causal effect of a nonrandomized public health policy from a study population to a target population. For instance, one may ask whether the causal effect of strict gun laws on crime in one state can be transportable to a different state. Methods are needed to account for differences in sociodemographic factors between the study and target populations and also account for spatiotemporal factors. For instance, the target state may be more rural, poorer, and have lower or higher crime rates. Further, the target state may be geographically closer to other states that have weaker gun laws, increasing the likelihood of traveling to another state to purchase guns. Novel methods for transporting causal effects that were originally estimated using, say, difference-in-differences, interrupted time series, or regression discontinuity designs under these complex settings would help investigators to better understand the causal effect of policy interventions in diverse settings. There is also room for developing methods for generalizing causal effects in these complex settings. For example, there may be interest in generalizing the causal effect of a successful drug prevention program in one school to all schools in the district. Careful thought would need to be given to the underlying framework and assumptions, including unmeasured confounding, interference, treatment heterogeneity, and potential mediation.

Quasi-experimental devices

An observational study investigates the effect of a treatment when the treatment is not randomized. A central concern is unmeasured differences between the treatment and control group, other than the treatment. After adjustments have been made for measured covariates in an observational study, an association between treatment and outcome is ambiguous: An association may be an effect caused by the treatment or a reflection of unmeasured differences between the treatment and control groups. Quasi-experimental devices enlarge the set of considered associations with the intention of reducing this ambiguity.

Classical quasi-experimental devices include pretreatment outcomes and multiple control groups (30). Recently, new quasi-experimental devices have been developed such as evidence factors, differential effects, and computerized construction of quasi-experiments (31). In much of the literature on classical quasi-experimental devices, which was developed by Donald Campbell and collaborators, statistical inference did not play a major role (30, 32). One potential future research direction is to work on incorporating statistical inference into the use of quasi-experimental devices. In most of the literature on classical and new quasi-experimental devices, the effect of one or at most a few treatments has been considered. Another important potential research direction is figuring out how current or novel quasi-experimental devices could help to make causal inferences about many treatments working together, as in a gene regulatory network.

DISCUSSION

Given the massive amount of data that are available today, along with machine learning tools, artificial intelligence tools, and high-performance computing, it is tempting for researchers to take more of a black box approach to answering scientific questions. However, we suggest it is essential for the causal inference community to continue to emphasize the importance of carefully laying out the problem, working with subject-matter experts to understand the data, and thinking carefully about the study design and plausibility of assumptions. We might not be known for flashy branding, but for data science to have the impact we think it can, principled causal inference approaches are critical.

A key consideration that we did not focus on in our list above is advances in estimation of causal estimands. Many of the new approaches we mentioned may benefit from semiparametric, nonparametric, or Bayesian estimation methods that focus on flexibility, multiple robustness, efficiency, and other properties.

We hope that the list above has convinced you that the field of causal inference has a very exciting future ahead. We are sure you can think of other areas of causal inference that would have been justified in claiming a spot in the top 10. We also anticipate that there are areas of research we do not yet know about but that will become critically important in the near future. We are excited to see where the field will go next!

ACKNOWLEDGMENTS

Author affiliations: Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, United States (Nandita Mitra); Department of Biostatistics and Epidemiology, Rutgers School of Public Health, Piscataway, New Jersey, United States (Jason Roy); and Department of Statistics, University of Pennsylvania, Philadelphia, Pennsylvania, United States (Dylan Small).

This work was funded by the National Institutes of Health (grants UL1TR003017 (J.R.) and 5R01AG065276-02 (D.S.)).

We thank members of the Center for Causal Inference for their responses to our survey.

The views expressed in this commentary are those of the authors and do not reflect those of the National Institutes of Health.

Conflict of interest: none declared.

REFERENCES

1. Pearl J. *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge University Press; 2013.
2. Hernan M, Robins J. *Causal Inference: What If?* Boca Raton, FL: Chapman & Hall; 2020.
3. Brumback B. *Fundamentals of Causal Inference: With R*. Boca Raton: Chapman & Hall/CRC; 2021.
4. Wang C, Dominici F, Parmigiani G, et al. Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models. *Biometrics.* 2015;71(3):654–665.
5. Ertefaie A, Asgharian M, Stephens DA. Variable selection in causal inference using a simultaneous penalization method. *J Causal Inference.* 2018;6(1):20170010.
6. Tang D, Kong D, Pan W, et al. Ultra-high dimensional variable selection for doubly robust causal inference [published online ahead of print January 19, 2022]. *Biometrics*. https://doi.org/10.1111/biom.13625.
7. Zhang Q. High-dimensional mediation analysis with applications to causal gene identification [published online ahead of print October 29, 2021]. *Stat Biosci*. https://doi.org/10.1007/s12561-021-09328-0.
8. Zigler CM. Invited commentary: the promise and pitfalls of causal inference with multivariate environmental exposures. *Am J Epidemiol*. 2021;190(12):2658–2661.
9. Nabi R, McNutt T, Shpitser I. Semiparametric causal sufficient dimension reduction of high dimensional treatments [preprint]. *arXiv*. 2020. https://doi.org/10.48550/arXiv.1710.06727. Accessed June 14, 2022.
10. Murphy SA. Optimal dynamic treatment regimes. *J R Stat Soc Series B Stat Methodology.* 2003;65(2):331–355.
11. Robins JM. *Optimal Structural Nested Models for Optimal Sequential Decisions*. In: Lin DY, Heagerty PJ, eds. *Proceedings of the Second Seattle Symposium in Biostatistics. Lecture Notes in Statistics*. New York, NY: Springer; 2004.
12. Caniglia EC, Sabin C, Robins JM, et al. When to monitor CD4 cell count and HIV RNA to reduce mortality and AIDS-defining illness in virologically suppressed HIV-positive persons on antiretroviral therapy in high-income countries: a prospective observational study. *J Acquir Immune Defic Syndr.* 2016;72(2):214–221.
13. Kreif N, Sofrygin O, Schmittdiel JA, et al. Exploiting nonsystematic covariate monitoring to broaden the scope of evidence about the causal effects of adaptive treatment strategies. *Biometrics.* 2021;77(1):329–342.
14. Microrandomized trials: an experimental design for developing just-in-time adaptive interventions. *Health Psychol.* 2015;34(0):1220–1228.
15. Luckett DJ, Laber EB, Kahkoska AR, et al. Estimating dynamic treatment regimes in mobile health using v-learning. *J Am Stat Assoc.* 2020;115(530):692–706.

16. Prentice RL, Langer RD, Stefanick ML, et al. Combined analysis of Women's Health Initiative observational and clinical trial data on postmenopausal hormone treatment and cardiovascular disease. *Am J Epidemiol.* 2006;163(7):589–599.
17. Kaizar EE. Estimating treatment effect via simple cross design synthesis. *Stat Med.* 2011;30(25):2986–3009.
18. Shortreed SM, Rutter CM, Cook AJ, et al. Improving pragmatic clinical trial design using real-world data. *Clin Trials.* 2019;16(3):273–282.
19. Neufeld A, Witten D. Discussion of Breiman's "two cultures": from two cultures to one. *Observational Stud.* 2021;7(1):171–174.
20. Kusner MJ, Loftus JR. The long road to fairer algorithms. *Nature.* 2020;578(7793):34–36.
21. Verbraeken J, Wolting M, Katzy J, et al. A survey on distributed machine learning. *ACM Comput Surv.* 2021;53(2):1–33.
22. Eberhardt F. Introduction to the foundations of causal discovery. *Int J Data Sci Anal.* 2017;3(2):81–91.
23. Glymour C, Zhang K, Spirtes P. Review of causal discovery methods based on graphical models. *Front Genet.* 2019;10:524.
24. Roberto CA, Lawman HG, LeVasseur MT, et al. Association of a beverage tax on sugar-sweetened and artificially sweetened beverages with changes in beverage prices and sales at chain retailers in a large urban setting. *JAMA.* 2019;321(18):1799–1810.
25. Beck B, Antonelli J, Piñeros G. Effects of New York City's neighborhood policing policy [published online ahead of print February 15, 2022]. *Police Q*. https://doi.org/10.1177/10986111211046991.
26. Reich BJ, Yang S, Guan Y, et al. A review of spatial causal inference methods for environmental and epidemiological applications. *arXiv*. 2020. https://doi.org/10.48550/arXiv.2007.02714. Accessed June 14, 2022.
27. Griffis H, Asch DA, Schwartz HA, et al. Using social media to track geographic variability in language about diabetes: infodemiology analysis. *JMIR Diabetes.* 2020;5(1):e14431.
28. Lee Y, Ogburn EL. Network dependence can Lead to spurious associations and invalid inference. *J Am Stat Assoc.* 2021;116(535):1060–1074.
29. Bareinboim E, Pearl J. A general algorithm for deciding transportability of experimental results. *J Causal Inference.* 2013;1(1):107–134.
30. Shadish W, Cook TD, Campbell DT. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin; 2001.
31. Rosenbaum PR. How to see more in observational studies: some new quasi-experimental devices. *Annu Rev Stat Appl.* 2015;2(1):21–48.
32. Stevenson JF, Campbell DT, Shadish W. Quasi-experimental designs. In: Atkinson P, Delamont S, Cernat A, et al., eds. *SAGE Research Methods Foundations*. Thousand Oaks, CA: SAGE Publications Ltd; 2020. https://dx.doi.org/10.4135/9781526421036914289. Accessed June 29, 2022.