BMC Medical Research Methodology

EDITORIAL Open Access

Causal inference and observational data

Ivan Olier[1*], Yiqiang Zhan[2], Xiaoyu Liang[3] and Victor Volovici[4]

Abstract

Observational studies using causal inference frameworks can provide a feasible alternative to randomized controlled trials. Advances in statistics, machine learning, and access to big data facilitate unraveling complex causal relationships from observational data across healthcare, social sciences, and other fields. However, challenges like evaluating models and bias amplification remain.

Main text

Billions of data records are generated every day, facilitating the discovery of knowledge. Particularly, medical, epidemiological, and social science research has significantly benefited from the vast amount of data available through sources such as medical records, easily attainable surveys, and social media platforms. This availability has led to a significant increase in the popularity of observational studies and meta-analyses as complementary approaches of randomized controlled trials (RCTs). RCTs are considered as the gold-standard study design for decision-making. However, conducting RCTs may not always be feasible due to ethical concerns, significant costs, or time limitations. Traditionally, outcomes from observational studies are considered of less value than RCTs, mainly because the former are vulnerable to confounding and bias issues. Recently, novel developments in statistics and machine learning (ML) are driven the development of causal inference in observational studies to the point of serving as a feasible substitute or complement for RCTs in decision-making [1]. Most statistical and ML methods are designed to establish an association map between input (factors) and output (target) variables. However, such association maps are unable to identify potential latent factors that influence both inputs and outputs, making their use limited to determining causal links. For instance, several studies reported a higher prevalence of lung cancer among coffee drinkers compared to non-drinkers. However, since many coffee drinkers also smoke, the observed association between coffee drinking and lung cancer is confounded by smoking, the true cause of the disease [2].

Causal inference from observational data finds application across various fields, with notable impact observed in domains such as healthcare, medicine, political and economic sciences, and social sciences. In healthcare and medical research, causal inference enables the identification of heterogeneous treatment effects and the formulation of personalized treatment strategies. By incorporating individual-level data, genetic information, and ML techniques, the field of personalized medicine benefits from enhanced causal inference methodologies [3]. The critical role of causal inference extends to policy evaluation and intervention assessment, where advancements in causal inference methods facilitate evidence-based decision-making by rigorously evaluating policy effectiveness, estimating causal impacts, and comprehending unintended consequences. Additionally, the

*Correspondence:
Ivan Olier
I.A.OlierCaparroso@ljmu.ac.uk
[1] Data Science Research Centre, Liverpool John Moores University, Liverpool, UK
[2] Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden
[3] Department of Epidemiology and Biostatistics, Michigan State University College of Human Medicine, East Lansing, MI 48824, USA
[4] Department of Neurosurgery, Center for Medical Decision Making, Erasmus MC, Rotterdam, The Netherlands

utilization of instrumental variables, regression discontinuity designs, and quasi-experimental approaches as methodological advancements further augment the understanding of complex social phenomena, policy impacts, and economic relationships [4, 5].

Broadly speaking, causal inference attempts to build data-driven models that can predict the effect of interventions on outcomes. Using observational data for causal inference is gaining momentum due to the confluence of factors such as the large amount of more complex and richer data and advanced techniques from statistics and ML. In general, two frameworks exist for causal inference in observational studies, which are not necessarily mutually exclusive: the structural causal model (SCM) framework and the potential outcome framework (POF). The SCM framework relies on deterministic, functional equations to construct directed acyclic graphs (DAGs) with variables as nodes and links as causal relationships and is particularly useful in identifying unknown causal and confounding variables while estimating the actual effect of a given treatment. On the other hand, the POF framework (also known as the counterfactual framework) examines outcomes that would have likely been observed had the treatment differed, representing the counterfactual or the missing outcome. Other frameworks such as instrumental variables, mediation analysis, and Bayesian networks are also noteworthy in causal inference research [6].

In recent years, there has been growing interest in combining multiple frameworks and approaches to improve causal inference. Integrating ideas from different frameworks can lead to more comprehensive and robust causal analyses. Additionally, the use of machine learning techniques and the exploration of new identification strategies are areas that hold promise for advancing causal inference research [7]. Analysis of observational studies could benefit from the best of two worlds. ML methods can help identify confounding variables, handle high-dimensional data, and improve prediction accuracy, while causal inference provides interpretability and causal understanding. Integrating these fields can lead to more powerful and robust causal inference models [8].

Causal inference research is a dynamic field that continues to evolve. Numerous real-world scenarios entail complex systems comprising multiple interacting variables. Advances in causal inference are instrumental in unraveling causal relationships in such systems. The availability of large-scale datasets presents both opportunities and challenges for causal inference. The development of scalable methods capable of efficiently handling large data sets while addressing biases, confounding, and selection effects constitute an active area of research. Furthermore, efforts are being made to devise methodologies for extracting causal relationships from unstructured data and integrating them with structured data, thereby enhancing the depth of insights and broadening the applicability of causal inference from observational data.

However, causal inference with observational data is not free of challenges. For instance, causal inference models are hard to evaluate. If a causal link is found, still there is no clear mechanism to assess whether the link is real or not. The performance of associative data-driven models can be assessed and compared easily since large data repositories are publicly available and widely used. However, this is not the case for causal inference, for which the lack of public benchmark data is one of the biggest problems it is encountered in their development. There is also a lack of comparisons to non-causal methods in the literature [9]. It is also inevitable to make untestable assumptions, which could also contribute to bias amplification and harm the external validity when compared to non-causal counterparts [10].

As the field continues to advance, interdisciplinary collaborations, methodological innovations, and the integration of emerging technologies will continue to expand the frontiers of causal inference and its applications in various domains. Nevertheless, challenges must be addressed for swift adoption in social and medical research.

Abbreviations

DAG	directed acyclic graphs
ML	machine learning
POF	potential outcome framework
RCT	randomized control trial
SCM	structural causal model

Authors' contributions

IO—conceived and drafted the Editorial. YZ, XL, VV revised the Editorial. All authors read and approved the final manuscript.

Funding

No funding was obtained for this editorial.

Data Availability

Not applicable.

Declarations

Competing interests

The authors of this editorial are Editorial Board Members of BMC Medical Research Methodology and Guest Editors of the Causal Inference and Observational Data collection.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Received: 21 September 2023 / Accepted: 6 October 2023
Published online: 11 October 2023

References

1. Hernán MA, Methods of Public Health Research — Strengthening Causal Inference from Observational Data. New England Journal of Medicine [Internet]. 2021 Oct 7 [cited 2023 May 23];385(15):1345–8. Available from: https://www.nejm.org/doi/full/https://doi.org/10.1056/NEJMp2113319.
2. Hemkens LG, Ewald H, Naudet F, Ladanie A, Shaw JG, Sajeev G, et al. Interpretation of epidemiologic studies very often lacked adequate consideration of confounding. J Clin Epidemiol. 2018;93:94–102.
3. Sanchez P, Voisey JP, Xia T, Watson HI, O'Neil AQ, Tsaftaris SA. Causal machine learning for healthcare and precision medicine. R Soc Open Sci. 2022;9(8).
4. Rohlfing I, Zuber CI. Check Your Truth Conditions!Clarifying the Relationship between Theories of Causation and Social Science Methods for Causal Inference. Sociol Methods Res [Internet]. 2021 Nov 1 [cited 2023 May 23];50(4):1623–59. Available from: https://journals.sagepub.com/doi/https://doi.org/10.1177/0049124119826156.
5. Varian HR, Proceedings of the National Academy of Sciences [Internet]. Causal inference in economics and marketing. 2016 Jul 5 [cited 2023 May 23];113(27):7310–5. Available from: https://www.pnas.org/doi/abs/https://doi.org/10.1073/pnas.1510479113.
6. Shi J, Norgeot B. Learning Causal Effects from Observational Data in Healthcare: a review and Summary. Front Med (Lausanne). 2022;9:864882.
7. Prosperi M, Guo Y, Sperrin M, Koopman JS, Min JS, He X, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nat Mach Intell. 2020;2(7):369–75.
8. Luo Y, Peng J, Ma J. When causal inference meets deep learning. Nature Machine Intelligence 2020 2:8 [Internet]. 2020 Aug 12 [cited 2023 May 23];2(8):426–7. Available from: https://www.nature.com/articles/s42256-020-0218-x.
9. Kaddour J, Lynch A, Liu Q, Kusner MJ, Silva R. Causal Machine Learning: A Survey and Open Problems. arXiv:220615475 [Internet]. 2022 Jun 30 [cited 2023 May 23]; Available from: http://arxiv.org/abs/2206.15475.
10. Hammerton G, Munafò MR. Causal inference with observational data: the need for triangulation of evidence. Psychol Med [Internet]. 2021 Mar 1 [cited 2023 May 23];51(4):563–78. Available from: https://www.cambridge.org/core/journals/psychological-medicine/article/causal-inference-with-observational-data-the-need-for-triangulation-of-evidence/AF5F7918753DF50F26B1D49561F0DF83.

Publisher's Note