carelon

carelonresearch.com

# Current applications of machine learning for causal inference in healthcare research using observational data

**Oluwadamilola Onasanya, MD, MPH[1]; Sarah Ruth Hoffman, PhD, MS, MPH[1]; Katherine Harris, PhD[1]; Ruth Dixon, PhD[1]; Michael Grabner, PhD[1]**

[1]Carelon Research, Wilmington DE

## Background

**Machine learning (ML): a broad concept with many applications.**

- ML refers to a family of statistical methods that generally focus on classification, ranking, and prediction with minimal human supervision.[1]
- ML approaches have the potential to facilitate causal inference using large, multidimensional real-world data (RWD).
- However, the implementation of ML-based approaches in answering causal pharmacoepidemiology questions can be conceptually and computationally complex.
- Researchers often ask the following questions:
  - *What are the potential applications of ML to my project?*
  - *What kinds of barriers are there to using RWD for causal inference in my study? How can ML help?*

## Objective

To visually illustrate the relationships between different types of barriers to causal inference in healthcare research and selected ML applications targeted at addressing these barriers, using an applied pharmacoepidemiological perspective.

## Methods

- We conducted a targeted review of published literature to identify RWD studies with ML applications for causal research.
- We classified the applications into three broad domains (presented via an infographic) and generated a list of illustrative case studies.

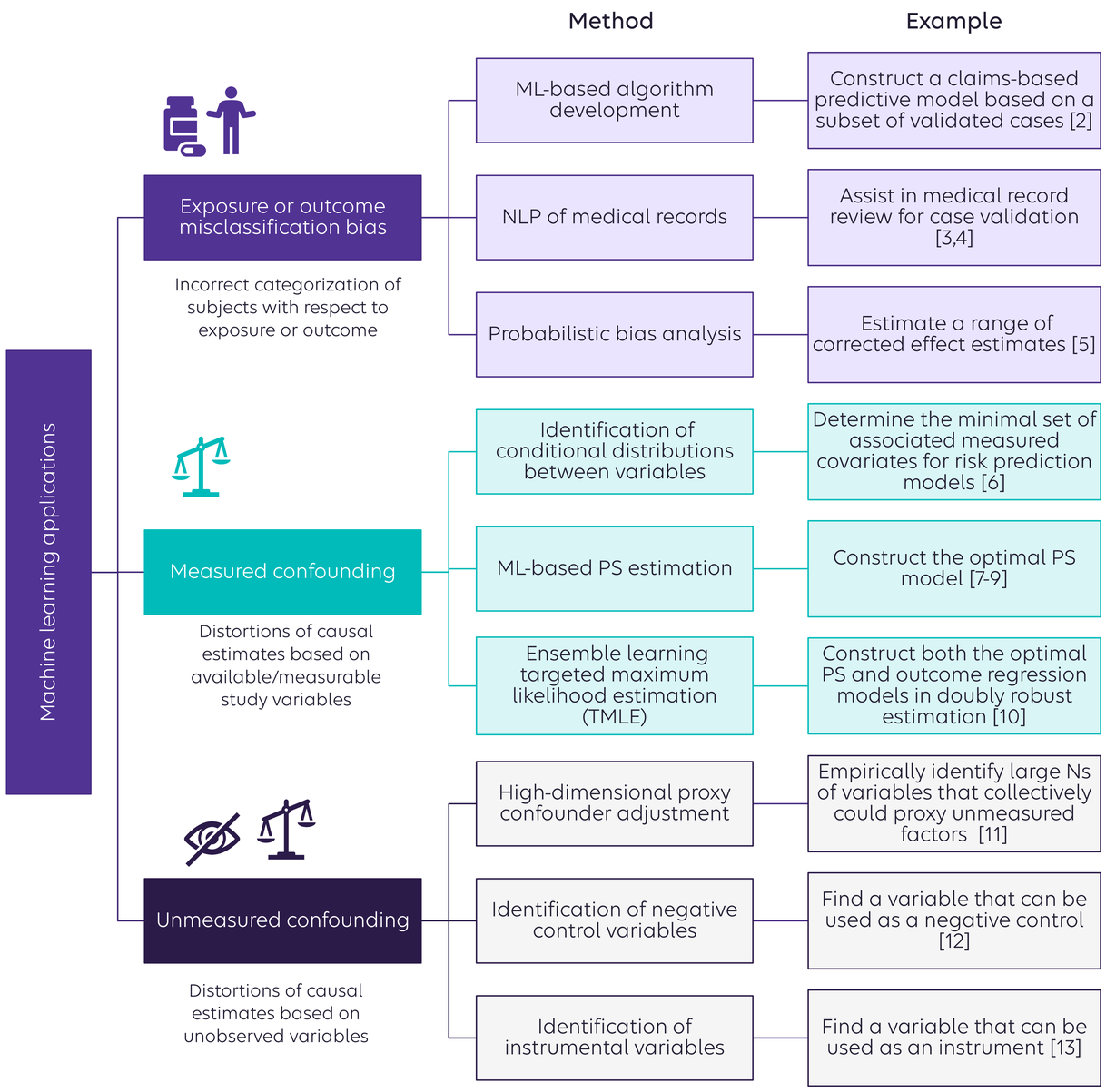
## Results

The identified ML applications were classified into three domains based on their potential to strengthen causal inference in pharmacoepidemiology studies.

## Funding and disclosures

No funding was received for the conduct of this study. All authors are employees of Carelon Research, which conducts health outcomes research with both internal and external funding, including a variety of private and public entities.

## Results (continued)




List of acronyms and terms:
High-dimensional proxy confounder adjustment: using large numbers of empirically identified features that collectively serve as proxies for unspecified or unmeasured confounders
Instrumental variable: a variable that affects the outcome of interest only through its effect on the exposure
Negative control variable: an alternative exposure (or outcome) variable whose relationship to the original outcome (or exposure) of interest does not use the same hypothesized causal pathway but is likely to involve the same sources of bias
NLP: natural language processing; a branch of AI concerned with the ability to support and manipulate human language
Probabilistic bias analysis: quantitative sensitivity analysis to assess the magnitude, direction, and uncertainty of bias using Monte Carlo techniques to repeatedly sample from bias parameter distributions
PS: propensity score; the conditional probability of being exposed given a set of covariates
Targeted maximum likelihood estimation: a doubly robust estimation method that includes a secondary targeting step that optimizes the bias-variance tradeoff for the parameter of interest

## Conclusions

- ML applications can strengthen causal research using real-world healthcare data. However, the range and complexity of applications limits wider use.
- To overcome this limitation, we provide a visual roadmap to relevant ML applications to help researchers quickly identify the appropriate tools given their specific research question.
- Note: The ML applications included in this poster are only a small fraction of all the potential utilities of ML for RWD research. This list can be expanded as new applications emerge and are developed.

## References

1. Padula WV, Kreif N, Vanness DJ, Adamson B, Rueda JD, Felizzi F, Jonsson P, IJzerman MJ, Butte A, Crown W. Machine learning methods in health economics and outcomes research—the PALISADE checklist: a good practices report of an ISPOR task force. Value in health. 2022 Jul 1;25(7):1063-80.
2. Beachler DC, de Luise C, Yin R, Gangemi K, Cochetti PT, Lanes S. Predictive model algorithms identifying early and advanced stage ER+/HER2- breast cancer in claims data. Pharmacoepidemiol Drug Saf. 2019 Feb;28(2):171-178. doi: 10.1002/pds.4681. Epub 2018 Nov 9. PMID: 30411431. [**Conducted by Carelon Research**]
3. Pfaff ER, Crosskey M, Morton K, Krishnamurthy A. Clinical Annotation Research Kit (CLARK): Computable Phenotyping Using Machine Learning. JMIR Med Inform. 2020 Jan 24;8(1):e16042. doi: 10.2196/16042. PMID: 32012059; PMCID: PMC7007592. [**Carelon Research co-author SRH was involved in this study**]
4. Hoffman SR, Pfaff ER, Nicholson WK. An Application of Machine Learning for the Refinement of an EHR-derived Cohort. Poster session presented at: 34th International Conference on Pharmacoepidemiology & Therapeutic Risk Management (ICPE). Prague, CZ. Aug 22-26, 2018. [**Carelon Research co-author SRH is lead author**]
5. No known implementations at this time. For more information on probabilistic bias analysis, see Lash, T.L., Fink, A.K., Fox, M.P. (2009). Probabilistic Bias Analysis. In: Applying Quantitative Bias Analysis to Epidemiologic Data. Statistics for Biology and Health. Springer, New York, NY. https://doi.org/10.1007/978-0-387-87959-8_8
6. Arora P, Boyne D, Slater JJ, Gupta A, Brenner DR, Druzdzel MJ. Bayesian Networks for Risk Prediction Using Real-World Data: A Tool for Precision Medicine. Value Health J Int Soc Pharmacoeconomics Outcomes Res. 2019;22(4):439-445. doi:10.1016/j.jval.2019.01.006
7. Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. J Clin Epidemiol. 2010 Aug;63(8):826-33. doi: 10.1016/j.jclinepi.2009.11.020. PMID: 20630332; PMCID: PMC2907172.
8. Belthangady C, Stedden W, Norgeot B. Minimizing bias in massive multi-arm observational studies with BCAUS: balancing covariates automatically using supervision. BMC Med Res Methodol. 2021 Sep 20;21(1):190. doi: 10.1186/s12874-021-01383-x. PMID: 34544367; PMCID: PMC8454087. [**This study was conducted by the parent company of Carelon Research.**]
9. Mai X, Teng C, Gao Y, Governor S, He X, Kalloo G, Hoffman SR, Mbiydzenyuy D, Beachler D. A pragmatic comparison of logistic regression vs. machine learning methods for propensity score estimation. Poster session presented at: 38th International Conference on Pharmacoepidemiology & Therapeutic Risk Management (ICPE); 2022 Aug 24-28; Copenhagen, Denmark. [**Conducted by Carelon Research**]
10. Kreif N, Tran L, Grieve R, De Stavola B, Tasker RC, Petersen M. Estimating the Comparative Effectiveness of Feeding Interventions in the Pediatric Intensive Care Unit: A Demonstration of Longitudinal Targeted Maximum Likelihood Estimation. Am J Epidemiol. 2017;186(12):1370-1379. doi:10.1093/aje/kwx213
11. Wyss R, Yanover C, El-Hay T, Bennett D, Platt RW, Zullo AR, Sari G, Wen X, Ye Y, Yuan H, Gokhale M, Patorno E, Lin KJ. Machine learning for improving high-dimensional proxy confounder adjustment in healthcare database studies: An overview of the current literature. Pharmacoepidemiol Drug Saf. 2022 Sep;31(9):932-943. doi: 10.1002/pds.5500. Epub 2022 Jul 5. PMID: 35729705; PMCID: PMC9541861.
12. Kummerfeld E, Lim J, Shi X. Data-driven Automated Negative Control Estimation (DANCE): Search for, Validation of, and Causal Inference with Negative Controls. Published online October 2, 2022. doi:10.48550/arXiv.2210.00528
13. Singh A, Hosanagar K, Gandhi A. Machine Learning Instrument Variables for Causal Inference. In: Proceedings of the 21st ACM Conference on Economics and Computation. EC '20. Association for Computing Machinery; 2020:835-836. doi:10.1145/3391403.3399466

Poster presented at ISPOR 2024, May 5-8, Atlanta, GA.

Dami Onasanya | Dami.Onasanya@carelon.com