IEA
International Epidemiological Association

International Journal of Epidemiology, 2020, 1763–1770
doi: 10.1093/ije/dyaa035
Advance Access Publication Date: 1 April 2020
Opinion

Opinion

Intersections of machine learning and epidemiological methods for health services research

Sherri Rose

Department of Health Care Policy, Harvard Medical School, 180 Longwood Ave, Boston, MA, 02115, USA. E-mail: rose@hcp.med.harvard.edu

Editorial decision 14 February 2020; Accepted 17 February 2020

Abstract

The field of health services research is broad and seeks to answer questions about the health care system. It is inherently interdisciplinary, and epidemiologists have made crucial contributions. Parametric regression techniques remain standard practice in health services research with machine learning techniques currently having low penetrance in comparison. However, studies in several prominent areas, including health care spending, outcomes and quality, have begun deploying machine learning tools for these applications. Nevertheless, major advances in epidemiological methods are also as yet underleveraged in health services research. This article summarizes the current state of machine learning in key areas of health services research, and discusses important future directions at the intersection of machine learning and epidemiological methods for health services research.

Key words: Machine learning, health services research, health quality, health outcomes

Key Messages

- Machine learning methods have been used less frequently in health services research, but are growing in health care spending, outcomes and quality.
- Many applied questions in health services research intersect with the methodological expertise of epidemiologists.
- Machine learning tools have promise for burgeoning areas in health services research, including methodology for difference-in-differences study designs.

Introduction

Health services research is a broad area focused on the health care system, including costs, quality, access to providers and services, and health outcomes following care. The field benefits from the interdisciplinary expertise of health policy scholars, clinicians, health economists, statisticians and public health researchers, as well as engagement from community members, policy makers and other stakeholders. Work in health services research is also published across an array of journals. While epidemiology is a distinct discipline studying the distributions, determinants and control of health events, there is an intersection with health services research, and epidemiologists have conducted key studies in health services research.

Data sources in health services research are not typically classical epidemiological cohorts, and often use health care billing claims, registry data, surveys or electronic health records. The latter three data sources are increasingly used in epidemiology, but health care billing claims, a staple of health services research, are less common in epidemiology. Each of these data sources has well-known advantages and disadvantages,[1–3] which will vary in importance depending on the research question.

For analysis, parametric regression techniques, rather than machine learning, are the standard in health services research. Machine learning methods aim to 'smooth' over the data, as traditional approaches also do, but they are often more flexible and may make fewer assumptions, typically operating in nonparametric or semiparametric models. Popular machine learning tools, such as tree-based techniques, neural networks and penalized regressions, have been used for classification questions and to identify high-risk individuals for health care interventions, but they have not been extensively integrated in health services research, especially not causal inference. The 'promise and perils' of these newer statistical learning tools for health services research have been discussed, with particular focus on the size of data repositories and sparsity of information.[4,5] This article highlights several areas where machine learning has begun to advance the field of health services research, and the role of epidemiological methods at this intersection.

Predicting health care spending

The financing of the health care system has many implications, including how health services for enrollees are provided and incentivized. Financing changes can also lead to improved health outcomes and access to care. For example, in *Better But Not Well*, authors Richard Frank and Sherry Glied discuss advances in mental health care over five decades that came not from new treatments but rather payment reforms and increased competition across providers, among other organizational changes.[6,7] Health care spending is studied from many perspectives, including spending levels or overall growth and by health condition.[8] The evaluation of new health payment policies is a central question in health services research and will be discussed in a later section on causality. Another impactful area is the risk adjustment of health plan payment formulas.

Plan payment risk adjustment aims to predict individual health spending Y using demographic and health condition variables $\boldsymbol{X}$ in order to reallocate funding according to the expected costs of a health plan's enrollees. This is an attempt to disincentivize avoiding high-cost enrollees, so that market competition is geared toward efficiency and quality.[2,9] Risk adjustment is used in many international health systems including in Belgium, Germany, The Netherlands, the USA and Israel. Epidemiologists will recognize this parametric regression problem:

$$E[Y|\boldsymbol{X}] = \beta \boldsymbol{X},$$

where Y is a bounded continuous outcome. This outcome Y might be transformed before the estimation procedure using the natural log or so-called 'top-coding' where all high-cost enrollees above a threshold dollar amount (e.g. \$250,000) are set to that threshold to improve performance with respect to specific metrics.[2] Prediction methods for health plan payment typically focus on parametric regression, with newer economics articles developing constrained regressions where the loss function is subject to certain restrictions.

Machine learning has thus far been applied only sparingly in the plan payment risk adjustment literature, and is often published in health services journals. The regression problem for machine learning is given as:

$$E[Y|\boldsymbol{X}] = f(\boldsymbol{X}),$$

where $f(\boldsymbol{X})$ is a flexible function of $\boldsymbol{X}$, which could include discovered features in $\boldsymbol{X}$. Three early papers in this space all considered regression trees, with one predicting payments for Medicare inpatient care,[10] another on Medicare psychiatric payments[11] and the last studying the addition of more complex interaction terms to predict total payments among commercially insured enrollees.[12] Tree-based methods create sequential splits of the data based on the provided covariates (or a subset of them) to yield groupings of observations that are highly homogeneous with respect to their outcome value. These techniques have become popular due to their ability to detect interactions and other non-linear relationships among the covariates. However, tree-based methods, including aggregation methods like random forests, may overfit to the training data even when using cross-validation. I refer interested readers to an accessible introductory machine learning book for further details on tree-based methods and other statistical learning techniques.[13]

Recent plan payment risk adjustment papers implemented ensembles of various learners to predict total payments[14,15] and mental health spending[16] among commercially insured enrollees, in addition to new work using regression trees to discover interaction terms, this time in the Dutch risk equalization formula.[17] Ensembles are a broad class of estimators that consider multiple algorithms to select either the single best algorithm (with respect to a particular criterion) or a weighted average of the algorithms. Tutorials on ensembles geared toward epidemiological audiences are available.[18,19]

Machine learning has also been deployed in the past 3 years in other health care spending application areas outside risk adjustment formulas. This includes demonstrating that health insurers can identify unprofitable enrollees in the unregulated United States Marketplace drug formularies, despite protections for pre-existing conditions.[20] Other studies predicted high-cost enrollees,[21] estimated cost-related health disparities[22] and predicted late-life spending.[23]

Whether the machine learning approaches for health spending discussed in this section appreciably improved on standard methods varied by study, and not all compared with a traditional approach. The practical utility of machine learning versus parametric regression is context-specific and may involve assessing the prediction functions along additional metrics not included in each article (e.g. if only R^2 was reported), as well as in external validation datasets. Evaluating algorithms using cross-validated metrics is good practice, but does not tell us how the prediction function will perform in data from subsequent years or if a prediction function created in Medicare fee-for-service enrollees is applicable to enrollees in private managed care Medicare plans.

Many ongoing practical estimation discussions surrounding health spending are centred on which variables should enter the algorithms, including the unintended consequences of incorporating social determinants of health,[2] using more comprehensive classification systems for categorizing health conditions[12,16] and the feasibility of integrating self-reported survey data at scale.[2,24] Other concerns focus on how to evaluate algorithms with respect to both statistical fit and fairness to marginalized groups,[22,25,26] and this is a major topic for future work. These considerations remain critical whether using parametric regression or machine learning. Epidemiologists' experience with prediction methods for continuous outcomes, evaluating prediction function performance along multiple dimensions, and the social contexts of using additional demographic information would augment the interdisciplinary teams building plan payment risk adjustment formulas and health care spending algorithms.

Predicting health outcomes and quality

Compared with health spending, there are many more examples of machine learning in health services research for the prediction of health outcomes and quality measures. A large portion, although not all, of these prediction functions consider binary outcomes, which can be written as:

$$\mathrm{logit}(P[Y = 1|\boldsymbol{X}]) = f(\boldsymbol{X}),$$

with $Y \in \{0, 1\}$. Mortality is assessed as a quality metric in some health services contexts, rather than exclusively as a health outcome. A number of recent papers have implemented machine learning to predict mortality, often among other outcomes, with respect to hospital performance.[27–30] One paper on the increasingly popular deep neural networks looked at mortality, readmission and length of stay, but these techniques had similar classification performance to regression methods when using a similar number of covariates.[31] Deep neural networks aim to define the strength of the associations between nodes across multiple constructed layers that form the 'network'. Like tree-based methods, deep neutral networks may find non-linear relationships in the data and are prone to overfitting, but may additionally discover novel features. Prediction of adverse events, adherence and rates of screening, testing and visits have also been explored as quality outcomes using machine learning.[30,32–35] Health outcomes studies have included predicting diabetes in claims data,[36] stroke risk,[37] obesity,[38] postoperative pain,[39] disease progression[40,41] and graft failure.[42] These health outcomes and quality studies were published across a spectrum of journals, most frequently in clinical journals.

Whereas health care quality is not a standard research question in epidemiology, health outcomes are commonly studied. Mortality prediction in particular is a frequent goal in epidemiological research, and epidemiologists' extensive knowledge, in developing risk scores and employing calibration and discrimination measures for binary outcomes, can enhance health outcomes and quality prediction work in health services research. Notably, machine learning for time-to-event outcomes in health services work is currently scarce. Most studies discretize mortality, length of stay and other outcomes such that they are binary. For a time-to-event outcome we have T the time to outcome Y, a censoring time C, $\tilde{T} = \min(T, C)$ the variable that defines which of T or C was observed earlier, and $\Delta = I(T = \tilde{T})$ an indicator for whether T was observed. The parameter of interest might be the conditional survival function $E[T > \theta|\boldsymbol{X}]$ (where θ is a time point threshold) or other choice. Machine learning applications for survival are understudied in both health services research and epidemiology. Survival research questions in health services research would benefit from collaborations with epidemiologists as both fields further integrate machine learning, given the penetrance of time-to-event epidemiological methods.

I close this section by highlighting that interpretability is a frequently raised query in considering machine learning for predicting health outcomes or quality. Performance metrics such as accuracy and calibration do not capture enough information to explain how the algorithm assigned outcomes. Because applications in health services research can have significant consequences, interpretability should be a priority.[43] Similarly, biases found in the underlying health data, including structural racism, can have massive implications if algorithms are deployed in practice.[44] Explainability and fairness are two features found in proposed social impact statements for algorithms.[45]

Causality, effect estimation and policy evaluation

Machine learning for causal inference is a newer area for most fields and has rarely been explored in health services research. Notable epidemiological methods development has occurred in this space, although infrequently applied. There are many causal contrasts that may be of interest, including the familiar average difference between the intervention and non-intervention groups:

$$\psi = E_X[E[Y|A = 1, \boldsymbol{X}] - E[Y|A = 0, \boldsymbol{X}]],$$

where $A \in \{0, 1\}$ is the intervention, which could be a treatment, exposure or policy. As is well known to epidemiologists, the validity of key causal assumptions in these studies is critical. In order to define our parameters causally, we must make a series of untestable assumptions: no unmeasured confounding, consistency and no interference between subjects, (as defined under the Neyman-Rubin causal framework), among other important assumptions. We can then write:

$$\psi = E_X[E[Y|A = 1, \boldsymbol{X}] - E[Y|A = 0, \boldsymbol{X}]] = E[Y^1 - Y^0],$$

where Y^1 and Y^0 are the counterfactual outcomes had everyone been set to receive the intervention and not receive the intervention, respectively. The use of machine learning in causal inference estimators does not obviate the need for thoughtful construction of an underlying causal model or magically remove data quality problems.[46]

A recent health services study (published in an epidemiology journal) estimated cancer mortality risk differences by emergency department presentation with double robust machine learning.[47] Double robust estimators will produce unbiased estimates for ψ if either the outcome regression, $E[Y|A, \boldsymbol{X}]$, or the probability of being in the intervention group given covariates, $P[A = 1|\boldsymbol{X}]$, is estimated consistently. By incorporating machine learning into double robust methods, $E[Y|A, \boldsymbol{X}]$ and $P[A = 1|\boldsymbol{X}]$ are estimated more flexibly and, especially when ensembles are used, minimal bias for ψ may be achieved in practice. A recent tutorial on these methods was published aimed at epidemiologists.[48] Although not yet frequently applied, causal inference incorporating machine learning has increased in the epidemiology literature,[49–55] with a number of studies using health care claims or electronic health record data. Issues particularly persistent in health services research with electronic health data that hinder causal inference, include missingness, misclassification and confounder selection. Variables may be collected irregularly, coding can vary by provider and facility and key confounders might be buried among hundreds of non-relevant variables. Variable selection techniques found in machine learning may aid in this last situation, but it is not guaranteed.

Comparative effectiveness research asks causal questions that consider the benefits and harms of health interventions and features substantial contributions from both health services scholars and epidemiologists.[56,57] Frequently, comparative effectiveness involves more than two treatments with more than one parameter or contrast of interest. For example, consider a treatment that has three binary levels representing different aortic valves: $\boldsymbol{A} = (A_1, A_2, A_3)$. Our parameters might be the three treatment-specific means:

$$\psi_1 = E_X[E[Y|A_1 = 1, \boldsymbol{X}]] = E[Y_1^1],$$

$$\psi_2 = E_X[E[Y|A_2 = 1, \boldsymbol{X}]] = E[Y_2^1],$$

$$\psi_3 = E_X[E[Y|A_3 = 1, \boldsymbol{X}]] = E\left[Y_3^1\right],$$

where Y_1^1, Y_2^1 and Y_3^1 are the counterfactual outcomes for having received each of the three valves, respectively. Machine learning has been examined in health services research for the comparative effectiveness of therapy, using tree-based methods in propensity score functions[58] and a continuous treatment on traumatic brain injury with ensembles,[59] as well as hip prosthesis on quality of life,[60] feeding interventions in the intensive care unit[61] and drug-eluting coronary artery stents,[62] all using double robust machine learning. In this last study, it was demonstrated empirically that the combination of double robust estimation and machine learning likely led to the isolation of individual stent effects. Comparative effectiveness parameters have parallels to variable importance studies where we create a ranked list of effect or association parameters, and are often found in genetic epidemiology (e.g. Winham *et al.*[63]), although typically without causal assumptions. Contemporary studies for variable importance of health conditions on health care spending[8] and ranking hospital quality based on excess mortality[64] both used double robust machine learning.

Policy evaluation is a major facet of health services research. One prevalent design to estimate the impact of new policies is a difference-in-differences approach. The policy intervention may be implemented at a particular level of geography with several other 'units' at the same geographical level selected to form a comparison group. Data from before the intervention and after intervention are required to estimate the parameter of interest. This parameter is often the difference between the intervention group in the post-intervention and pre-intervention periods minus the difference between the comparison group in both time periods, hence the name 'difference-in-differences'. The difference-in-differences parameter can be written causally and recognized as an average treatment effect among the treated:

$$\psi = E\left[Y^1_{POST} - Y^0_{POST}A = 1\right],$$

where the subscript *POST* represents the post-intervention time period. It is important to stress that causal interpretation of this parameter requires thoughtful consideration of the required causal assumptions.[65] Machine learning research for difference-in-differences studies is extremely limited.[66] However, recent work in the creation of so-called synthetic comparison groups (i.e. weighted averages among units) has incorporated machine learning (e.g. Amjad *et al.*[67]). Both parameter estimation and the construction of suitable comparison groups are vital areas for future machine learning work in policy evaluation.

Looking forward

Health services research as a field is less flashy than many domains publicizing dramatic advances using 'artificial intelligence' methods, but this is not to say that careful, reasoned machine-learning work will not lead to progress in improving health care costs, quality, access, outcomes and additional areas not discussed in this piece. A focus on the external validation, generalizability and reproducibility of research results is crucial for health services findings to lead to actual successes in practice. Additionally, any time we are using data not collected for research purposes—common in health services research—we must pay extra attention to identifying the underlying processes that generated the data, which is aided by working with a diverse interdisciplinary research team.

The expertise of epidemiologists will be valuable in these teams as use of machine learning increases in health services research. This article described several areas where epidemiological methods can contribute, including causal inference, techniques for time-to-event outcomes and the inclusion of social determinants of health. However, working across disciplines is challenging. Epidemiologists may need to learn new machine learning concepts and jargon in order to communicate across these barriers, as well as additional programming languages (e.g. R and Python). Knowledge of the intricacies of the health care system is also paramount to avoid spurious results—from minutiae like changes in billing code standards to broad issues such as physician behaviour.

Growth areas for machine learning in health services research will likely encompass study designs and parameters frequently seen in the economics and policy literature, including difference-in-differences approaches discussed earlier, and instrumental variables[68] studies. Experimental studies incorporating machine learning to reduce variance is another area.[69] Unsupervised statistical learning methods, such as clustering, have been employed to group observations as stand-alone research questions for some time. Clustering has also been integrated into evaluations in order to study impact by groups (e.g. Lee *et al.*[70]). One consequence of the increase in the number of available variables in electronic health data resources is that evaluations conditional on algorithm-defined groups might become more common. This may be especially true for precision medicine applications and studies of treatment effect heterogeneity, two additional topics where epidemiologists have substantial insights. Machine learning also has promise for contributing to a learning health care system (e.g. Deeny and Steventon[71]). Last, data linkages across disparate sources, including imaging, wearable technology, streaming public data (e.g. Twitter) and unstructured data (e.g. text fields in electronic health records), are exciting but in need of continued, rigorous vetting.

Health services research often examines or seeks to inform policy, and machine-learning studies have strong potential to contribute to comprehensive evidence synthesis for such policy changes. Although far from comprehensive in its scope, this article has summarized key intersections between machine learning and epidemiological methods for health services research. There is great promise for progress in the future, made even more likely by further leveraging the expertise of epidemiologists.

Funding

This work was supported by the National Institutes of Health through an NIH Director's New Innovator Award (DP2-MD012722).

Conflict of interest

None declared.

References

1. Tyree PT, Lind BK, Lafferty WE. Challenges of using medical insurance claims data for utilization analysis. *Am J Med Qual* 2006;**21**:269–75.
2. Ellis R, Martins B, Rose S. Risk adjustment for health plan payment. In: McGuire T, van Kleef R (eds). *Risk Adjustment, Risk Sharing and Premium Regulation in Health Insurance Markets: Theory and Practice*. New York, NY: Elsevier, 2018.
3. Haneuse SJ, Shortreed SM. On the use of electronic health records. In: Gatsonis C, Morton S (eds). *Methods in Comparative Effectiveness Research*. New York, NY: Chapman and Hall/CRC, 2017.
4. Crown WH. Potential application of machine learning in health outcomes research and some statistical cautions. *Value Health* 2015;**18**:137–40.
5. Frakt AB, Pizer SD. The promise and perils of big data in health care. *Am J Manag Care* 2016;**22**:98–9.
6. Frank RG, Glied SA. *Better but Not Well: Mental Health Policy in the United States since 1950*. Baltimore, MD: JHU Press, 2006.
7. McGuire TG. Achieving mental health care parity might require changes in payments and competition. *Health Aff (Milwood)* 2016;**35**:1029–35.
8. Rose S. Robust machine learning variable importance analyses of medical conditions for health care spending. *Health Serv Res* 2018;**53**:3836–54.
9. Iezzoni L. *Risk Adjustment for Measuring Healthcare Outcomes*. Chicago, IL: Health Administration Press, 1997.
10. Relles D, Ridgeway G, Carter G. Data mining and the implementation of a prospective payment system for inpatient rehabilitation. *Health Serv Outcomes Res Methodol* 2002;**3**:247–66.
11. Drozd EM, Cromwell J, Gage B, Maier J, Greenwald LM, Goldman HH. Patient casemix classification for Medicare psychiatric prospective payment. *Am J Psychiatry* 2006;**163**: 724–32.
12. Robinson JW. Regression tree boosting to adjust health care cost predictions for diagnostic mix. *Health Serv Res* 2008;**43**: 755–72.
13. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning with Applications in R*. New York, NY: Springer, 2013.
14. Rose S. A machine learning framework for plan payment risk adjustment. *Health Serv Res* 2016;**51**:2358–74.
15. Sungchul P, Anirban B. Alternative evaluation metrics for risk adjustment methods. *Health Econ* 2018;**27**:984–1010.
16. Shrestha A, Bergquist SL, Montz E, Rose S. Mental health risk adjustment with clinical categories and machine learning. *Health Serv Res* 2018;**53**:3189–206.
17. van Veen S, van Kleef RC, van de Ven W, van Vliet R. Exploring the predictive power of interaction terms in a sophisticated risk equalization model using regression trees. *Health Econ* 2018;**27**:e1–12.
18. Rose S. Mortality risk score prediction in an elderly population using machine learning. *Am J Epidemiol* 2013;**177**:443–52.
19. Naimi A, Balzer L. Stacked generalization: An introduction to super learning. *Eur J Epidemiol* 2018;**33**:459–64.
20. Rose S, Bergquist SL, Layton TJ. Computational health economics for identification of unprofitable health care enrollees. *Biostatistics* 2017;**18**:682–94.
21. Tamang S, Milstein A, Sorensen HT *et al*. Predicting patient 'cost blooms' in Denmark: a longitudinal population-based study. *BMJ Open* 2017;**7**:e011580.
22. Bergquist SL, Layton TJ, McGuire TG, Rose S. Data transformations to improve the performance of health plan payment methods. *J Health Econ* 2019;**66**:195–207.
23. Einav L, Finkelstein A, Mullainathan S, Obermeyer Z. Predictive modeling of U.S. health care spending in late life. *Science* 2018;**360**:1462–65.
24. Rose S, Zaslavsky AM, McWilliams JM. Variation in accountable care organization spending and sensitivity to risk adjustment: implications for benchmarking. *Health Aff (Milwood)* 2016;**35**:440–48.
25. Rose S, McGuire TG. Limitations of p-values and R-squared for stepwise regression building: a fairness demonstration in health policy risk adjustment. *Am Stat* 2019;**73**:152–56.
26. Zink A, Rose S. Fair regression for health care spending. *Biometrics* 2019;doi: 10.1111/biom.13206.
27. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. *Lancet Respir Med* 2015;**3**:42–52.
28. Mansoor H, Elgendy IY, Segal R, Bavry AA, Bian J. Risk prediction model for in-hospital mortality in women with ST-elevation myocardial infarction: a machine learning approach. *Heart Lung* 2017;**46**:405–11.
29. DeCenso B, Duber HC, Flaxman AD, Murphy SM, Hanlon M. Improving hospital performance rankings using discrete patient diagnoses for risk adjustment of outcomes. *Health Serv Res* 2018;**53**:974–90.
30. Bihorac A, Ozrazgat-Baslanti T, Ebadi A *et al*. MySurgeryRisk: development and validation of a machine-learning risk algorithm for major complications and death after surgery. *Ann Surg* 2019;**269**:652–62.
31. Rajkomar A, Oren E, Chen K *et al*. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med* 2018;**1**:18.
32. Hubbard RA, Zhu W, Balch S, Onega T, Fenton JJ. Identification of abnormal screening mammogram interpretation using Medicare claims data. *Health Serv Res* 2015;**50**:290–304.
33. Franklin JM, Shrank WH, Lii J *et al*. Observing versus predicting: initial patterns of filling predict long-term adherence more accurately than high-dimensional modeling techniques. *Health Serv Res* 2016;**51**:220–39.
34. Chirikov VV, Shaya FT, Onukwugha E, Mullins CD, dosReis S, Howell CD. Tree-based claims algorithm for measuring pretreatment quality of care in Medicare disabled hepatitis C patients. *Med Care* 2017;**55**:e104–12.
35. Larney S, Hickman M, Fiellin DA *et al*. Using routinely collected data to understand and predict adverse outcomes in

opioid agonist treatment: Protocol for the Opioid Agonist Treatment Safety (OATS) Study. *BMJ Open* 2018;**8**:e025204.

36. Razavian N, Blecker S, Schmidt AM, Smith-McLallen A, Nigam S, Sontag D. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data* 2015;**3**:277–87.
37. Mullainathan S, Obermeyer Z. Does machine learning automate moral hazard and error? *Am Econ Rev* 2017;**107**:476–80.
38. Dugan TM, Mukhopadhyay S, Carroll A, Downs S. Machine learning techniques for prediction of early childhood obesity. *Appl Clin Inform* 2015;**6**:506–20.
39. Tighe PJ, Harle CA, Hurley RW, Aytug H, Boezaart AP, Fillingim RB. Teaching a machine to feel postoperative pain: combining high-dimensional clinical data with machine learning algorithms to forecast acute postoperative pain. *Pain Med* 2015;**16**:1386–401.
40. Konerman MA, Zhang Y, Zhu J, Higgins PD, Lok AS, Waljee AK. Improvement of predictive models of risk of disease progression in chronic hepatitis C by incorporating longitudinal data. *Hepatology* 2015;**61**:1832–41.
41. Konerman MA, Lu D, Zhang Y *et al.* Assessing risk of fibrosis progression and liver-related clinical outcomes among patients with both early stage and advanced chronic hepatitis C. *PLoS One* 2017;**12**:e0187344.
42. Lau L, Kankanige Y, Rubinstein B *et al.* Machine-learning algorithms predict graft failure after liver transplantation. *Transplantation* 2017;**101**:e125–32.
43. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *ArXiv*, arXiv: 1702.08608, 28 February 2017, preprint: not peer reviewed.
44. Chen J, Asch S. Machine learning and prediction in medicine—beyond the peak of inflated expectations. *N Engl J Med* 2017;**376**:2507–09.
45. Diakopoulos N, Friedler S, Arenas M *et al.* Principles for accountable algorithms and a social impact statement for algorithms. fatml.org/resources/principles-for-accountable-algorithms (16 July 2019, date last accessed).
46. Petersen M, van der Laan M. Causal models and learning from data: Integrating causal modeling and statistical estimation. *Epidemiology* 2014;**25**:418–26.
47. Luque-Fernandez MA, Belot A, Valeri L, Cerulli G, Maringe C, Rachet B. Data-adaptive estimation for double-robust methods in population-based cancer epidemiology: risk differences for lung cancer mortality by emergency presentation. *Am J Epidemiol* 2018;**187**:871–78.
48. Schuler M, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies. *Am J Epidemiol* 2017;**185**:65–73.
49. 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;**63**:826–33.
50. Padula AM, Mortimer K, Hubbard A, Lurmann F, Jerrett M, Tager IB. Exposure to traffic-related air pollution during pregnancy and term low birth weight: estimation of causal associations in a semiparametric model. *Am J Epidemiol* 2012;**176**:815–24.
51. Franklin JM, Eddings W, Glynn RJ, Schneeweiss S. Regularized regression versus the high-dimensional propensity score for confounding adjustment in secondary database analyses. *Am J Epidemiol* 2015;**182**:651–59.
52. Pirracchio R, Petersen ML, van der Laan M. Improving propensity score estimators' robustness to model misspecification using super learner. *Am J Epidemiol* 2015;**181**:108–19.
53. Schneeweiss S, Eddings W, Glynn RJ, Patorno E, Rassen J, Franklin JM. Variable selection for confounding adjustment in high-dimensional covariate spaces when analyzing healthcare databases. *Epidemiology* 2017;**28**:237–48.
54. Karim ME, Pang M, Platt RW. Can we train machine learning methods to outperform the high-dimensional propensity score algorithm? *Epidemiology* 2018;**29**:191–98.
55. Wyss R, Schneeweiss S, van der Laan M, Lendle SD, Ju C, Franklin JM. Using super learner prediction modeling to improve high-dimensional propensity score estimation. *Epidemiology* 2018;**29**:96–106.
56. Gatsonis C, Morton S (eds). *Methods in Comparative Effectiveness Research*. New York, NY: Chapman and Hall/CRC, 2017.
57. Levy A, Sobolev B (eds). *Comparative Effectiveness Research in Health Services*. New York, NY: Springer, 2017.
58. Watkins S, Jonsson-Funk M, Brookhart MA, Rosenberg SA, O'Shea TM, Daniels J. An empirical comparison of tree-based methods for propensity score estimation. *Health Serv Res* 2013;**48**:1798–817.
59. Kreif N, Grieve R, Díaz I, Harrison D. Evaluation of the effect of a continuous treatment: a machine learning approach with an application to treatment for traumatic brain injury. *Health Econ* 2015;**24**:1213–28.
60. Kreif N, Gruber S, Radice R, Grieve R, Sekhon JS. Evaluating treatment effectiveness under model misspecification: a comparison of targeted maximum likelihood estimation with bias-corrected matching. *Stat Methods Med Res* 2016;**25**:2315–36.
61. 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**:1370–79.
62. Rose S, Normand SL. Double robust estimation for multiple unordered treatments and clustered observations: evaluating drug-eluting coronary artery stents. *Biometrics* 2019;**75**:289–96.
63. Winham SJ, Jenkins GD, Biernacka JM. Modeling X chromosome data using random forests: conquering sex bias. *Genet Epidemiol* 2016;**40**:123–32.
64. Spertus JV, Normand SLT, Wolf R, Cioffi M, Lovett A, Rose S. Assessing hospital performance after percutaneous coronary intervention using big data. *Circ Cardiovasc Qual Outcomes* 2016;**9**:659–69.
65. Zeldow B, Hatfield L. Difference-in-differences. diff.healthpolicydatascience.org (16 July 2019, date last accessed).
66. Weber AM, van der Laan MJ, Petersen ML. Assumption trade-offs when choosing identification strategies for pre-post treatment effect estimation: an illustration of a community-based intervention in Madagascar. *J Causal Inference* 2015;**3**:109–30.

67. Amjad M, Shah D, Shen D. Robust synthetic control. *J Mach Learn Res* 2018;**19**:1–51.
68. Athey S, Tibshirani J, Wager S. Generalized random forests. *Ann Statist* 2019;**47**:1148–78.
69. Jones D, Molitor D, Reif J. *What Do Workplace Wellness Programs Do? Evidence From the Illinois Workplace Wellness Study.* Cambridge, MA: National Bureau of Economic Research, Working Paper #24229, 2018.
70. Lee I, Monahan S, Serban N, Griffin PM, Tomar SL. Estimating the cost savings of preventive dental services delivered to Medicaid-enrolled children in six southeastern states. *Health Serv Res* 2018;**53**:3592–616.
71. Deeny SR, Steventon A. Making sense of shadows: priorities for creating a learning healthcare system based on routinely collected data. *BMJ Qual Saf* 2015;**24**:505–15.

International Journal of Epidemiology, 2020, 1770–1773
doi: 10.1093/ije/dyaa242
Advance Access Publication Date: 28 November 2020

Commentary: Towards machine learning-enabled epidemiology

Louisa R Jorm*

Centre for Big Data Research in Health, Faculty of Medicine, University of New South Wales, Sydney, Australia

*Corresponding author. Centre for Big Data Research in Health, Faculty of Medicine, UNSW Sydney, Kensington, NSW 2052, Australia. E-mail: l.jorm@unsw.edu.au

Editorial decision 30 October 2020

Epidemiology, data science and machine learning

Since first emerging as a discipline in the 1990s, data science has become a critical area of workforce skills shortage.[1] Although data science has no agreed definition, it is centred in multidisciplinary and interdisciplinary approaches to extracting knowledge or insights from data for use in a broad range of applications.[1] The role of the epidemiologist in the health and medical domain aligns strongly with a common definition of a data scientist as someone who 'combines domain-specific expertise with analytic skills to extract knowledge from data to drive action'.[2] However, most training programmes in epidemiology do not teach the primary skills that healthcare organizations seek in data scientists, which include machine learning (ML) and the open-source programming languages R and Python.[3] Indeed, a course in data science was a mandatory component of only 18% of epidemiology programmes offered by the top 20-ranked public health schools in the USA in 2019.[4]

There has been considerable discussion within the statistical community regarding the relationship between statistics, data science and ML,[5] emphasising the need to ensure that statisticians have the necessary skills in computation. Engineering and computer science graduates are seen as currently better equipped than statisticians to contribute as data scientists.[6] Forging new approaches that bring together ML and statistical communities and mindsets is presented as a solution to addressing challenges inherent in the application of ML to big datasets including selection bias, measurement error, quantifying uncertainty, and interpretability.[7]

It is still early days for similar discussions among epidemiologists. However, commentators argue that whereas epidemiologists do not necessarily need to learn coding at the expense of core epidemiological skills,[4,8] or become experts in ML,[4] they do need a foundational knowledge of data science techniques to equip them to work in the large interdisciplinary teams that will make big discoveries in science. The pervasive use of closed-source programming languages (e.g. SAS, Stata) is cited as being a barrier to the integration of ML techniques in epidemiology.[9]

Advancing epidemiologists' awareness of machine learning

The burgeoning use of ML across all aspect of health and medicine creates an imperative for epidemiologists to be at the very least 'ML-aware'. Three papers in this issue of the *International Journal of Epidemiology* serve to advance this cause. All three focus on the unique power of ML methods for prediction.