International Journal of Epidemiology, 2020, 2058–2064
doi: 10.1093/ije/dyz132
Advance Access Publication Date: 11 July 2019
Education Corner

Education Corner

Reflection on modern methods: when worlds collide—prediction, machine learning and causal inference

Tony Blakely,[1]* John Lynch,[2] Koen Simons,[1] Rebecca Bentley[1] and Sherri Rose[3]

[1]Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Victoria, Australia, [2]School of Public Health, University of Adelaide, Adelaide, South Australia, Australia and [3]Department of Health Care Policy, Harvard Medical School, Boston, Massachusetts, USA

*Corresponding author. Melbourne School of Population and Global Health, University of Melbourne, 207 Bouverie Street, Carlton, Melbourne, VIC 3010, Australia. E-mail: ablakely@unimelb.edu.au

Editorial decision 31 May 2019; Accepted 11 June 2019

Abstract

Causal inference requires theory and prior knowledge to structure analyses, and is not usually thought of as an arena for the application of prediction modelling. However, contemporary causal inference methods, premised on counterfactual or potential outcomes approaches, often include processing steps before the final estimation step. The purposes of this paper are: (i) to overview the recent emergence of prediction underpinning steps in contemporary causal inference methods as a useful perspective on contemporary causal inference methods, and (ii) explore the role of machine learning (as one approach to 'best prediction') in causal inference. Causal inference methods covered include propensity scores, inverse probability of treatment weights (IPTWs), G computation and targeted maximum likelihood estimation (TMLE). Machine learning has been used more for propensity scores and TMLE, and there is potential for increased use in G computation and estimation of IPTWs.

Key words: Machine learning, causal inference, prediction, potential outcomes

Key Messages

- Contemporary causal inference methods in epidemiology often include pre-final estimation steps predicting propensity scores or potential outcomes.
- Machine learning algorithms aim to 'learn' or predict outputs (e.g. exposed/unexposed, outcomes) from inputs (covariates) in a new sample, having been first trained on a training dataset that contains both inputs and labelled outputs.
- Machine learning is starting to be used in pre-final steps of contemporary causal inference methods and there is potential for increased use.

Introduction

In epidemiology, prediction and causal modelling are usually considered as different worlds.

Prediction modelling uses information 'within-the-data' to create a model that most accurately predicts something, a characteristic or outcome of interest. One common and clinically useful class of prediction modelling identifies who is likely to get maximal absolute treatment benefit from therapies (proven elsewhere to be effective with—say—randomized controlled trials). For example, who will likely gain the most from statin treatment to prevent cardiovascular disease (CVD).[1] The variables that perform well in that prediction may not be causal for disease, e.g. high-density lipoproteins (HDLs) strongly predict CVD risk, but HDL itself is not causal of CVD as shown in Mendelian randomization studies.[2]

Causal modelling rigorously tests hypotheses generated from theory and content knowledge external to the data with explicit attention to key assumptions such as consistency and exchangeability (i.e. no confounding). The use of directed acyclic graphs (DAGs; this and some other specific terms in the text are defined in the glossary in Table 1)[5,6] is current best practice for bringing prior knowledge, theory and a formally defined data structure to any analysis seeking to identify causal effects. In this paradigm, only those variables that are confounders (or on back door paths) should be adjusted for in commonly used analytical methods ranging from stratification through to multivariable regression modelling. It is incorrect to adjust for variables that are not on back door paths, and in particular it is incorrect to adjust for intermediaries (those variables on the causal pathway from exposure to outcome, or front door

Table 1. A glossary of terms and concepts used in this paper

Term	Definition and/or concept
Back door path	A non-causal path in a DAG from exposure to outcome that has an arrow coming into the exposure. If there is no collider on the back door path, it is open and requires blocking by conditioning for one of more variables on the path.
Collider	A variable or node on a path in a DAG from exposure to outcome that has both arrows pointing into it.
Confounder	A. Has three properties: 1. Must be associated with the exposure in the source population 2. Must be an extraneous risk factor for the disease: - Need not be actual cause of disease, but must be surrogate of cause - Must be risk factor among non-exposed (in the source population) 3. Must not be affected by (common cause of) the exposure or (common cause of) disease. In particular, it cannot be an intermediary. B. A variable on an open back door path in a DAG.
Directed acyclic graph (DAG)	A causal diagram where all arrows are directed and represent causal effects on one variable on another, and is acyclic in that one cannot return to where one started via directed arrows.
Ensemble learning	A technique using multiple algorithms (and could include traditional regression methods) that combines them to improve estimates and predictive performance. Types of ensemble models include random forests, bagging, boosting and stacking (or super learner).
Front door path	A causal path in a DAG from exposure to outcome that has an arrow going out of exposure, and arrow into the outcome, and no colliders.
G computation	Is a 'maximum likelihood substitution estimator of the G-formula. . .. [and is] equivalent to using the marginal distribution of the covariates as the standard in standardization, a familiar class of procedures in epidemiology'. (Snowden *et al.*[3])
Inverse probability of treatment weights (IPTWs)	The inverse of the propensity score (PS). IPTWs are commonly used to estimate parameters defined by marginal structural models for a time-varying exposure or treatment as well as in cross-sectional studies.
Machine learning	Algorithms that aim to 'learn' or predict outputs (exposed/unexposed, treated/untreated) from inputs (covariates) in a new sample, having been first trained on a training dataset that contains both inputs and labelled outputs.
Propensity score	The probability of being exposed or treated, using an equation based on confounders.
Targeted maximum likelihood estimation (TMLE)	'Is a doubly robust maximum-likelihood-based approach that includes a secondary "targeting" step that optimizes the bias-variance trade-off for the target parameter'. (Schuler and Rose[4]) For the average treatment effect (ATE), it involves both outcome modelling (akin to G computation) and exposure modelling (akin to PS, but more to optimize the bias variance trade-off – hence 'targeted').

paths) when estimating the effect, and it is incorrect to adjust for colliders (i.e. those variables inducing a selection bias if adjusted for).

Causal inference methods and best-prediction modelling have become less distinct in recent years due to the development of causal inference methods (often premised on a potential outcomes approach[7] or structural causal models[8]) that harness predictive estimation in pre-final estimation steps. For example, the prediction of inverse probability of treatment weights (IPTWs) as a step before their use in a weighted estimator. Rapid developments in computer science, especially machine learning algorithms that allow for selection of main terms, interactions and non-linear relationships to better fit the observed data, accentuate the potential for sophisticated and automated predictive estimation steps in analytical strategies that aim to make epidemiological causal inference.[9–11]

The purpose of this paper is not to review in depth machine learning or causal inference. (Regarding machine learning in epidemiology, the reader is instead directed to: accompanying papers in this issue of *IJE* and other reviews of machine learning from an epidemiological perspective.[10,12]) Rather, the purposes of this paper are: (i) to overview the recent emergence of prediction underpinning steps in contemporary causal inference methods as a useful perspective on contemporary causal inference methods, and (ii) explore the role of machine learning (as one approach to 'best prediction') in causal inference.

Unless stated otherwise, we focus on the average treatment effect (ATE) in the population as a whole, or an effect that would be given by comparing the whole population had they been exposed to the whole population had they been unexposed. Table 2 provides supporting information to the sections below.

Predicting exposures: propensity scores

Propensity scores (PS) reduce information on multiple confounding covariates into one value: the propensity to be exposed or treated,[13] i.e. $\Pr(X = 1|Z)$ for a binary exposure X and a vector of covariates Z. The generation of a PS is a pre-effect estimation step, with the propensity scores used in the final outcome model by way of matching exposed and unexposed subjects with similar PS or using the PS as inverse weights. Consistent estimation of the PS strengthens internal validity of subsequent outcome modelling, by adjusting for confounding. Within the confines of selecting the Z covariates to model the PS (i.e. they are confounders; and they are not exogenous predictors of just X), the best specification of covariates Z and model specification is flexible. Put another way, we are agnostic to what transformations (e.g. log, cubic splines, etc.) and interactions of (possibly transformed) covariates Z are used, and how these Z covariates are used to predict X (e.g. regression, decision trees, classification algorithms). We might just want the most accurate prediction or PS that also optimally balances confounders between the exposed and unexposed. To do so, it may be more efficient to use machine learning algorithms, rather than manual, time consuming user-specification with trial and error of various algorithms.

Indeed, many of the early epidemiological applications of machine learning in causal inference have been to calculate PS. The earliest example (according to[14]) is a simulation study by Setoguchi *et al.*[15] comparing recursive partitioning and neural networks with logistic regression. The two machine learning methods arguably outperformed logistic regression, but the gains (reductions in bias) were small and sometimes at the expense of less precision (i.e. wider standard errors) of the final X−Y association determined in the outcome regression using PS matching. Examples of machine learning generated PS have followed since with some gains in confounding control.[14,16–19] Recently, machine and ensemble learning methods have been applied to not only best prediction of exposure, but optimal selection and modelling of covariates in the propensity score algorithm based on optimizing the balance of confounding covariates between the exposed and unexposed.[16,20]

Predicting weights for exposures: inverse probability of treatment weights (IPTWs)

The PS (as stated above) can also be used to weight analyses with 1/PS for the exposed (or treated), and 1/(1-PS) for the unexposed (or untreated). In a simple cohort study with no repeated measures of exposure and covariates, this inverse weighting by PS will adjust for baseline confounding and may provide the same benefit as matching, regressing or stratifying on the PS. However, IPTWs can also be used with repeated measures data where variables may be intermediaries for the association of exposure at one point in time with the outcome, but also confounders of the association for the (time varying) exposure at future points in time with the outcome. IPTWs are commonly used to estimate parameters defined by marginal structural models.[21] As with PS, (user-specified) logistic regression is the most common method to calculate IPTWs, but also as with PS the IPTWs have no causal interpretation themselves—making them natural quantities for estimation with machine learning.

For example, Bentley *et al.*[22] aimed to estimate the impact of cumulative exposure to social housing, and transitions in and out of social housing, on mental health. They used ensemble learning (combining three types of 'base learners': logistic regression with cubic b-splines; a gradient

Table 2. When can pure prediction help causal inference in epidemiology?

Method	Pre-final causal effect estimation that involves prediction	Final step	Prediction guidelines	Could machine learning assist?
1. Standard	No	Regression $E[Y\|X, Z] = f(X, Z)$		Probably not, as even if only confounders and exposure included it may be hard to extract a meaningful effect size per unit change in exposure.
2. Propensity scores	Yes – predicting exposure (pr(X|Z)).	Matching by pr(X|Z) or adjusting by pr(X|Z) and then estimating the effect of interest.	Include confounders of X→Y association, plus perhaps predictors of Y. Optimize both prediction of X, and selection of (residually) confounding variables.	Yes. We do not interpret coefficients in the propensity score prediction function, so best prediction of propensity score is desired.
3. Inverse probability of treatment weights (IPTWs)	Yes – predicting exposure and constructing inverse probability of treatment weights at each time step, t ($IPTW_t$).	Inverse weighting by pr(X|Z) for exposed and 1−pr(X|Z) for unexposed; weighting by product of $IPTW_t$ across time.	Include time invariant and varying confounders up to time t for each $IPTW_t$	Yes. We do not interpret coefficients in the propensity score prediction functions that create the IPTWs.
4. G computation	Yes – predicting potential outcomes, e.g. $E[Y^{X=x}]$ for counterfactual intervention on exposure X; $E[Y^{M=m}]$ for counterfactual intervention on mediator M.	Analysis of the predicted outcomes (not the observed outcomes) under both exposure and non-exposure for all individuals.	Use exposure and covariates that meet standard confounder properties.	Yes. We do not interpret coefficients or effect sizes in the equation predicting $E[Y^{X=x}]$, $E[Y^{M=m}]$, etc.
6. Targeted maximum likelihood estimation (TMLE)	Yes – predicting both the outcome and the exposure.	Following the targeted 'update' step that incorporates information from the propensity score function to reduce bias, analyses compare the predicted outcomes under both exposure and non-exposure.	Include all potential confounders in both prediction functions.	Yes, both for predicting the outcome and predicting the exposure. Including an ensemble that contains methods that perform variable selection can help bring about meaningful reduction of the number of potential confounders.

X, exposure; Y, outcome; Z, confounding covariates; M, mediators.

boosting machine; and a conditional inference forest). Compared with standard logistic regression estimation of IPTWs, the ensemble learner's weights were superior in two respects: a narrower distribution of IPTWs; and better balance of covariates between exposed and unexposed (although it was still not ideal). Exposure-outcome estimates using ensemble learning IPTWs were notably different to using standard logistic regression IPTWs, albeit with overlapping confidence intervals. We are aware of only a few other examples of machine learning to generate IPTWs in marginal structural models published in epidemiological journals (e.g.[23]); this seems a fertile area to incorporate machine learning into epidemiological causal inference.

Predicting outcomes: G computation and other methods

Following the adage that potential outcomes are the ultimate missing data problem[24], epidemiologists are increasingly explicitly estimating individual outcome status had they been counterfactually (un)exposed[3,25] or experienced differing levels of mediating risk factors.[26–28] This estimation, or prediction, of potential outcomes (and potential mediators) is, again, a pre-final effect estimation step. We are not seeking to interpret coefficients or other parameters used in these prediction algorithms. Rather, we are predicting outcome values for all individuals, then using this expanded dataset to directly estimate causal effects of interest, be that the marginal ATE or effect sizes within strata of the data (e.g. by sex). For example, we may estimate the average of every individual's difference in outcome under exposure and unexposed (at least one being counterfactual).[3,25,29] Given that, for this simple example at least, we could use a standard regression model to estimate an effect size for the exposure–outcome association to undertake prediction of potential outcomes, why bother? First, it decouples the estimation of the causal effects per se from the estimation of all other parameters required[3]—a conceptual advantage. Second, in the presence of heterogeneity of the exposure–outcome association across levels of covariates (i.e. effect modification), one can both estimate the marginal average treatment effect for the population averaged across this heterogeneity, as well as conditional within subsets of the population. Third, with predicted outcomes it is simple to visualize the outcome risks or rates by exposure in graphs, and to calculate effect measures on both absolute and relative scales—enabling, in our experience at least, simpler reporting for readers and end-users (e.g.[27]).

Again, we are predicting potential outcomes as a pre-final estimation step, and the final step may be as simple as averaging the individual differences in potential outcomes across individuals. An early example of using machine learning to predict potential outcomes found that it outperformed standard methods when the outcome model was non-linear and non-additive (i.e. the true predictive equation had quadratic terms and many interactions of predictors).[30] The above is a form of G computation, which when the underlying functional form is simple may be better estimated with standard regression modelling (i.e. parametric G computation[31]). It can also be used for research questions that have a longitudinal nature, such as 'what is the effect of an intervention programme that increases tobacco cessation in middle age on later development of cardiovascular disease?'. For this question, we want to allow for the fact that in the absence of the intervention smokers are still likely to quit at older ages for 'business as usual' (BAU) reasons. Such estimators require sequential estimation steps, often using parametric regressions, and extensive calibration of the prediction equations in BAU before estimating the counterfactual intervention.[31] As 'big data' access improves, such approaches to answer policy-relevant questions are likely to increase. Machine learning to undertake these predictions at each time-step, within the confines of only using covariates that are on back door paths at any point in time (i.e. not intermediaries or colliders), seems a fertile opportunity to exploit machine learning. Westreich *et al.* (2015)[25] state that these types of estimators 'can be made more robust to model misspecification through machine-learning techniques'.[14] However, examples to date are sparse (e.g.[30]), although we anticipate this to be a growth area for epidemiology and related fields.

Blended exposure and outcome modelling: doubly robust, targeted maximum likelihood estimation methods

Doubly robust methods[32] for the ATE have an exposure model (e.g. PS) in addition to the outcome regression model that includes covariate adjustment. The beauty of doubly robust methods, and from where their name derives, is that only one of the exposure model or the outcome model needs to be correctly specified (or more broadly, estimated consistently) for the final parameter estimators to be unbiased. If both are estimated consistently,[32] the estimator will also be asymptotically efficient. The procedure is increasingly used. For example, in the paper described above by Bentley *et al.* (2018)[22] that used ensemble learning to construct IPTWs in a study of the association of social housing with mental health, they also adjusted 'again' for some of the covariates used in the IPTW calculation in the outcome regression.

The most common use of the double robust method with machine learning prediction for causal inference is in the targeted maximum likelihood estimator (TMLE).[4,33]

For a simple ATE, it involves outcome prediction (just as in the G computation estimator) and additionally includes an updating or targeting step that incorporates information from the PS. This updating step optimizes the bias-variance trade-off for the parameter of interest rather the overall outcome regression distribution. Machine learning can be used for both the outcome and exposure modelling. Tutorials aimed at epidemiologists have been published for TMLEs using machine learning with both continuous and binary outcomes, and include R code for TMLE implementation as well as G computation and propensity score estimators.[4,34] TMLEs possess many favourable statistical properties beyond their double robustness, including having a loss-based principle for dealing with multiple solutions.[33]

Thus, the potential gains from double robust machine learning are 2-fold: we not only have two opportunities to obtain unbiased estimation of our final estimator, but we are more likely to obtain at least one consistently estimated outcome or exposure regression by considering machine learning (and specifically ensembles).[35] Imagine the scenario where a simple main-terms regression is misspecified for the outcome and exposure regressions. In this case, if we included that main-terms regression in an ensemble of other learners that are better able to search the covariate space, we have protected our final estimates from this bias as the ensemble assigns lower or zero weight to the misspecified regression(s). Simulation studies find that under a range of circumstances, including with large, collinear covariate sets, double robust analyses may be more accurate (less systematic error or bias) than either outcome or exposure methods used in isolation.[32,36] Conversely, it is true that machine learning may not always be an improvement over traditional approaches used to estimate the outcome and exposure regressions. For example, if the underlying functional form is well estimated by a main-terms regression, a double robust machine learning estimator that considers many learners (including parametric regression) will still be unbiased, but may have slightly larger confidence intervals (see Schuler and Rose[4] for examples).

What else and what next?

The recruitment of machine learning into causal inference methods is largely about achieving exchangeability—or accounting for confounding—be that through propensity scores, weighting, or potential outcome prediction. Machine learning has been (and is likely to be increasingly) used to identify effect heterogeneity,[35] with recent methodological work (for example) demonstrating how random forests combined with the potential outcomes approach can robustly detect and estimate heterogeneity of treatment effects across multiple covariates considered simultaneously.[37] We anticipate increasing cross-over from computer science—including machine learning methods—into epidemiology for methods to address measurement error and missing data. Methods such as regression calibration,[38] quantitative bias analysis[39] and multiple over-imputation[40] exist, albeit arguably under-utilized in epidemiology. Machine learning may offer some assistance for mismeasurement of confounders, if only by being able to include more variables in prediction modelling steps; even if variables are mismeasured or unmeasured, if they are correlated then including more of them may help block back door paths by correlation.[41]

Conclusion

Advances in causal inference methods and the emergence of big, complex, longitudinal data as well as data science, will profit from incorporating methods such as machine learning into epidemiological causal inference. The different worlds of prediction and causal modelling inference have blurred. As with any new method, machine learning is no panacea—and may not always gain as much in accuracy and precision for the resources invested as epidemiologists might expect, but that 'cost' will decrease as the methods become more familiar. We argue that thinking about the pre-final estimation steps in causal inference—prediction that can be aided by machine learning—offers a useful conceptual approach to deploy potential outcomes thinking in epidemiology.

Funding

T.B. was supported by a Health Research Council of New Zealand Programme Grant (16/443). R.B. was funded by Australian Research Council (ARC) Future Fellowships (FT150100131). J.W.L. was supported by an NHMRC Centre of Research Excellence (GNT1099422). S.R. was supported by an NIH Director's New Innovator Award (DP2-MD012722).

Acknowledgements

We acknowledge assistance from Lizzie Korevaar and Rob Mahar with literature searching.

Conflict of interest: None declared.

References

1. Pylypchuk R, Wells S, Kerr A. Cardiovascular disease risk prediction equations in 400000 primary care patients in New Zealand: a derivation and validation study. *Lancet* 2018;**391**:1897–907.
2. Voight BF, Peloso GM, Orho-Melander M *et al*. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet* 2012;**380**:572–80.

3. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *Am J Epidemiol* 2011;**173**:731–38.
4. Schuler MS, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies. *Am J Epidemiol* 2017;**185**:65–73.
5. Glymour M, Greenland S. Causal diagrams. In: Rothman K, Greenland S, Lash T (eds). *Modern Epidemiology*. 3rd edn. Philadelphia: Lippincott WIlliams & Wilkins, 2008, pp. 183–212.
6. Greenland S, Pearl J, Robins J. Causal diagrams for epidemiologic research. *Epidemiology* 1999;**10**:37–48.
7. Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annu Rev Public Health* 2000;**21**:121–45.
8. Pearl J. *Causality*, 2nd edn. Cambridge: Cambridge University Press, 2009.
9. Mooney SJ, Pejaver V. Big data in public health: terminology, machine learning, and privacy. *Annu Rev Public Health* 2018; **39**:95–112.
10. Keil AP, Edwards JK. You are smarter than you think: (super) machine learning in context. *Eur J Epidemiol* 2018;**33**:437–40.
11. Rose S. Mortality risk score prediction in an elderly population using machine learning. *Am J Epidemiol* 2013;**177**:443–52.
12. Naimi AI, Balzer LB. Stacked generalization: an introduction to super learning. *Eur J Epidemiol* 2018;**33**:459–64.
13. Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;**70**: 41–55.
14. Westreich D, Lessler J, Funk MJ. Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *J Clin Epidemiol* 2010;**63**:826–33.
15. Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiol Drug Saf* 2008;**17**:546–55.
16. Karim ME, Pang M, Platt RW. Can we train machine learning methods to outperform the high-dimensional propensity score algorithm? *Epidemiol* 2018;**29**:191–98.
17. 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.
18. Setodji CM, McCaffrey DF, Burgette LF, Almirall D, Griffin BA. The right tool for the job: choosing between covariate-balancing and generalized boosted model propensity scores. *Epidemiology* 2017;**28**:802–11.
19. 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.
20. Pirracchio R, Carone M. The Balance Super Learner: a robust adaptation of the Super Learner to improve estimation of the average treatment effect in the treated based on propensity score matching. *Stat Methods Med Res* 2018;**27**:2504–18.
21. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiol* 2000;**11**:550–60.
22. Bentley R, Baker E, Simons K, Simpson JA, Blakely T. The impact of social housing on mental health: longitudinal analyses using marginal structural models and machine learning-generated weights. *Int J Epidemiol* 2018;**47**:1414–22.
23. Gruber S, Logan RW, Jarrin I, Monge S, Hernan MA. Ensemble learning of inverse probability weights for marginal structural modeling in large observational datasets. *Stat Med* 2015;**34**: 106–17.
24. Holland PW. Statistics and causal inference. *J Am Stat Assoc* 1986;**81**:945–60.
25. Westreich D, Edwards JK, Cole SR, Platt RW, Mumford SL, Schisterman EF. Imputation approaches for potential outcomes in causal inference. *Int J Epidemiol* 2015;**44**:1731–37.
26. 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.
27. Blakely T, Disney G, Valeri L *et al*. Socioeconomic and tobacco mediation of ethnic inequalities in mortality over time: repeated census-mortality cohort studies, 1981 to 2011. *Epidemiology* 2018;**29**:506–16.
28. Chittleborough CR, Mittinty MN, Lawlor DA, Lynch JW. Effects of simulated interventions to improve school entry academic skills on socioeconomic inequalities in educational achievement. *Child Dev* 2014;**85**:2247–62.
29. Naimi AI, Cole SR, Kennedy EH. An introduction to G methods. *Int J Epidemiol* 2017;**46**:756–62.
30. Austin PC. Using Ensemble-based methods for directly estimating causal effects: an investigation of tree-based G-computation. *Multivariate Behav Res* 2012;**47**:115–35.
31. Keil AP, Edwards JK, Richardson DB, Naimi AI, Cole SR. The parametric g-formula for time-to-event data: intuition and a worked example. *Epidemiology* 2014;**25**:889–97.
32. Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly robust estimation of causal effects. *Am J Epidemiol* 2011;**173**:761–67.
33. van der Laan M, Rose S. *Targeted Learning: Causal Inference for Observational for Experimental Data*. New York: Springer, 2011.
34. Luque-Fernandez MA, Schomaker M, Rachet B, Schnitzer ME. Targeted maximum likelihood estimation for a binary treatment: a tutorial. *Stat Med* 2018;**37**:2530–46.
35. Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. *Proc Natl Acad Sci USA* 2016;**113**:7353–60.
36. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics* 2005;**61**:962–73.
37. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. *J Am Stat Assoc* 2018; **113**:1228–42.
38. Buonaccorsi J. *Measurement Error: Models, Methods, and Applications*. New York: Chapman & Hall/CRC, 2010.
39. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. *Int J Epidemiol* 2014;**43**:1969–85.
40. Blackwell M, Honaker J, King G. A unified approach to measurement error and missing data: overview and applications. *Sociol Methods Res* 2015;**46**:303–41.
41. Fewell Z, Davey Smith G, Sterne J. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. *Am J Epidemiol* 2007;**166**:646–55.