Vol. 190, No. 8
https://doi.org/10.1093/aje/kwab047
Advance Access publication:
March 6, 2021

Thirteen Questions About Using Machine Learning in Causal Research (You Won't Believe the Answer to Number 10!)

Stephen J. Mooney*, Alexander P. Keil, and Daniel J. Westreich

* Correspondence to Dr. Stephen J. Mooney, Department of Epidemiology, School of Public Health, University of Washington, 3980 15th Avenue NE, Seattle, WA 98195 (e-mail: sjm2186@uw.edu).

Initially submitted October 22, 2019; accepted for publication October 16, 2020.

Machine learning is gaining prominence in the health sciences, where much of its use has focused on data-driven prediction. However, machine learning can also be embedded within causal analyses, potentially reducing biases arising from model misspecification. Using a question-and-answer format, we provide an introduction and orientation for epidemiologists interested in using machine learning but concerned about potential bias or loss of rigor due to use of "black box" models. We conclude with sample software code that may lower the barrier to entry to using these techniques.

causal inference; double-robustness; epidemiologic methods; inverse probability weighting; machine learning; propensity score; targeted maximum likelihood estimation

Abbreviations: LASSO, least absolute shrinkage and selection operator; TMLE, targeted maximum likelihood estimation.

Editor's note: *An invited commentary on this article appears on page 1483.*

Machine learning has recently gained prominence within epidemiology (1–4) and other fields incorporating data analysis (5–8). Much of this attention is given to a branch of machine learning called *supervised learning*, which uses algorithms to uncover patterns within the data in hand with the goal of making accurate predictions for other data (9, 10). At first blush, such work has a different purpose than causal epidemiologic research, which uses assumptions derived from background knowledge about confounding, mediation, and effect modification (11) to make inference from the data in hand. However, in some settings, estimates of causal effects have been improved by selectively incorporating machine-learning techniques within a broader estimation framework (12–14). While "machine learning" is a vast field that defies easy summary, in this commentary we ponder answers to common questions epidemiologists have about how machine learning might facilitate causal research.

THIRTEEN QUESTIONS ABOUT MACHINE LEARNING

What is "machine learning," really?

Machine learning, broadly, is a discipline focused on deriving algorithms that yield optimal predictions or decisions from data. In our experience, epidemiologists typically use the term "machine learning" to refer to supervised learning, the subfield whose methods treat 1 data column as the "supervisor" and then use algorithms to find patterns in other data columns to predict it (3). For example, estimating regression functions, as we do when fitting parametric models, is a form of supervised learning in which the dependent variable or outcome is the supervisor.

However, whereas regression functions are typically fitted via investigator-specified models that are (ideally) chosen prior to seeing the data, the techniques that are usually considered machine learning use *automated* algorithms that identify which aspects of the data best predict the outcome using techniques that include but are not limited to automated variable selection and flexible variable smoothing. Forward stepwise regression, an algorithmic approach to progressively adding covariates into a regression model until

the addition of covariates no longer improves model fit, is a form of machine learning.

Technically, this notion of identifying the "best" model is about identifying the model that minimizes a loss function, which is a function that assigns a cost to each observation in a training data set. The trained parameters of the algorithm can then be used to compute an expected outcome given observed predictors, either in the training data or in a different set of data (9). Common loss functions include the squared difference between the predicted variable and the observed target variable (the loss function for ordinary least squares regression), the difference in log likelihood (the loss function for logistic regression), and the absolute difference between prediction and observation.

Doesn't machine learning answer predictive questions but not causal questions?

Much of the present excitement around machine learning focuses on predictive modeling, and machine learning *is* useful for questions that don't focus on cause-effect relationships, including measurement, risk assessment, prognosis, and imputation. For example, machine learning has recently been used to measure neighborhood conditions from Google Street View imagery (Google, Inc., Mountain View, California) (15), to recover unobserved fetal weights (16), and to infer gestational age from data available at birth (17). Better measures and more accurate imputation remove potential biases in causal effect estimates.

It is less widely understood that if careful consideration is paid to caveats that we detail below, machine learning can also be embedded directly within the process of causal effect estimation.

How can I incorporate machine learning into a causal analysis?

First, note that causal model misspecification (e.g., improperly adjusting for a mediator or collider in a model used to estimate a causal effect) and statistical model misspecification (e.g., "mean model misspecification," such as falsely specifying that 2 nonlinearly related variables have a linear relationship, or "distributional misspecification," such as modeling a log-normally distributed variable with a normal distribution) are distinct errors (18). In observational epidemiology, we typically use causal models (such as causal directed acyclic graphs (19) or single world intervention graphs (20)) to decide how to specify and interpret statistical models. Causal models typically cannot be inferred from data alone and must be generated from background knowledge (21). There exists a branch of machine learning, known as causal discovery, that focuses on the conditions and assumptions under which causal models can be inferred from temporal ordering of changes in data alone (22, 23), though the granularity and scale of data currently required for these approaches render them unlikely to be useful to most epidemiologists in the near term (and so are not the focus of this paper).

By contrast, a statistical model—the distributions and covariances of variables in multidimensional space—*can* be (imperfectly) inferred from data. Often, machine learning is used not to infer a full statistical model but rather to estimate the conditional expectation of the outcome given predictors (sometimes called a "mean model") with which to predict outcomes in new data.

In summary, while causal models describe hypothesized processes and can identify variables as exposures, mediators, confounders, etc., there are many possible regression models compatible with a single directed acyclic graph. We therefore reference these causal models when we specify statistical models from which we estimate statistical parameters. When causal identifiability criteria hold, these statistical estimates can be interpreted as estimates of causal effect (24).

When we provide these same confounder variables to machine-learning algorithms, the algorithms—with limited additional input from the investigator—can similarly fit a mean model (sometimes called a nuisance function or nuisance model) that can be used to account for confounding. For example, instead of an investigator deciding whether or not to include age in a statistical model as 1) a linear term, 2) a categorical variable, 3) a polynomial, or 4) a restricted cubic spline, a machine-learning algorithm provided with age in its variable set can decide for itself how best to relate age to exposure and outcome, potentially selecting empirically from a wide variety of choices or dropping age entirely. Given typical loss functions, these data-driven decisions typically reduce concerns about statistical model misspecification.

Wait a second! If machine-learning algorithms can drop variables, what if they drop my exposure??!

Whereas epidemiologists thinking causally specify a contrast in a particular treatment or exposure, which elevates the role of a specific variable and orients model choices around this variable, machine-learning algorithms that optimize a bias-variance tradeoff in reference to a loss function (see "What is machine learning, really?" for technical details about a loss function) may neglect exposures of interest in favor of more predictive variables. For example, the least absolute shrinkage and selection operator (LASSO) regression technique—a common machine-learning approach—selects a subset of potential predictors that together minimize absolute prediction error (25). In the presence of strong confounding, a LASSO regression could select a strong confounder rather than your exposure of interest as an outcome predictor. If we used this LASSO regression to estimate counterfactual outcomes as in a conventional outcome regression model, we might well see no effect of the exposure (i.e., it would "shrink" all the way to 0).

However, there are several ways to mitigate this concern. First, many machine-learning approaches, such as generalized additive models, do not perform variable selection and thus are not subject to this phenomenon. Second, with a large sample size, we can fit machine-learning algorithms separately within each stratum of the exposure and predict outcomes separately in each arm, thus bypassing this problem entirely. Finally, we could select a method that is not vulnerable to this phenomenon. For example, if we

use machine-learning approaches to estimate the score for a matched propensity score analysis, the observed exposure serves as the "outcome" of the machine-learning algorithm, and so the exposure cannot be removed by the algorithm (though this approach requires caution regarding variable selection and the bias-variance tradeoff, as detailed below) (26, 27).

What is this bias-variance tradeoff thing?

The bias-variance tradeoff is a key idea in both statistics and machine learning and can be confusing to epidemiologists because its name incorporates familiar words but uses them to refer to related but subtly different notions. The "bias" of the bias-variance tradeoff refers to a difference between the mean value of model predictions computed from repeated samples and the true value of the parameter being estimated. The "variance" characterizes the sensitivity of the predictions to which observations from an underlying population happen to be in the data set to which the model was fitted. In general, model-building choices that let a model fit the observed data more closely (e.g., allowing there to be more knots in a spline model, which results in less bias in the observed data) result in models whose final form is more influenced by the specific data to which the model is fitted—that is, are higher-variance. Thus, too much bias results in underfitted models and too much variance results in overfitted models. Both can make models predict badly.

A number of machine-learning algorithms have "tuning parameters"—inputs to fitting a given model that optimize the bias-variance tradeoff—often selected using cross-validation. Cross-validation is a process in which an analyst sets aside a portion of their data (the "test set") and fits a model to the rest of the data (the "training set"). By testing how well models that they fit to the training set with different tuning parameters predict within the test set, they can empirically identify a tuning parameter that balances bias and variance to minimize error. The most common form of cross-validation, K-fold cross validation, repeats this partitioning K times to cover the full data set.

Cool. So machine learning can help me pick the covariates to put into my model?

It is again helpful to distinguish between causal models and statistical models. Machine learning cannot identify the causal structure giving rise to data: Algorithms cannot in general distinguish confounders from mediators, nor can they determine whether key confounders are unmeasured or whether a variable is a collider. If your data are missing variables, you would need to meet the causal identifiability criteria; no amount of machine learning can overcome this.

But given a set of potential (noncollider, nonmediator) confounding variables to include in a model, machine learning *can* help select a subset of those variables that reduce confounding bias while avoiding model instability due to multicollinearity and avoiding loss of precision due to excessive variance from including too many predictor variables. However, be cautious: Automated variable selection can cause problems. For example, when using machine learning to compute propensity scores, model selection will favor instruments of the exposure, which can induce finite sample bias or magnify uncontrolled confounding from other sources (26, 28).

Still, if you have enough background knowledge to distinguish between weak predictors and nonconfounders and between pure instruments and true confounders, machine learning can improve variable selection.

How do I know when machine learning will provide a more correct estimate than a conventional regression model?

Parametric regression—wherein prespecified covariates are modeled with a (frequently linear) relationship with the outcome (or log odds of a dichotomous outcome)—will only accurately estimate a parameter interpretable as a causal effect to the extent that the statistical model is correct (21). Thus, even with an accurate causal model, statistical misspecification can lead to incomplete confounding control.

Machine-learning algorithms typically allow more flexible modeling forms or different modeling assumptions (6). For example, tree-based methods such as classification and regression trees or random forests specifically do not require linearity assumptions. We caution that flexibility is no panacea; for example, tree-based approaches may be poorly suited to predicting in contexts where discontinuities (that is, sharp categorizations) are implausible (29).

Machine learning often identifies covariate patterns that predict exposure (or outcome) better than an a priori model (often including linearity and homogeneity assumptions) would. In many instances, the resulting model is better able to reduce statistical model misspecification and consequent residual confounding. Several empirical simulations suggest that propensity scores generated using machine learning improve model performance in the presence of 3 or more product terms in the data-generating model, relative to a naive linear model (1, 30).

An additional concern regarding data-driven model selection is that frequentist statistical inference assumes models are correct and all uncertainty derives from data. As a result, confidence intervals from an analysis that naively embeds machine learning may fail to incorporate model uncertainty and so may be too narrow (13). Fortunately, this problem may be mitigated by sample splitting, embedding machine learning in doubly robust estimation techniques such as targeted maximum likelihood estimation (TMLE), or both (12, 13, 31). Whereas the latter approach relies on deriving variance estimates from the influence curve—a function describing how the estimator would change if the observed data were slightly different (32)—the former approach accounts for model uncertainty by design, because model selection is performed using distinct data. More broadly, embedding machine learning in doubly robust estimators may also minimize bias resulting from the "curse of dimensionality," wherein flexible models fitted with many predictors require much more data to be statistically consistent, and thus could result in more biased estimates in practice (33).

How do I interpret the estimate I get if I use machine learning to estimate a causal effect?

The causal interpretation of any effect you estimate is determined both by the identifiability assumptions you use to link your causal model to data and by the statistical procedure you use to estimate the causal parameter (24). For example, if you perform an inverse-probability–weighted analysis, the statistical parameter you estimate will estimate an average treatment effect when causal identification conditions hold. This will be true regardless of whether the propensity scores are computed using standard parametric models or flexible machine-learning algorithms.

However, the numerical result will typically differ if the flexible algorithm chooses a model that differs from the standard parametric model. Assuming that the more flexible model better approximates the true data-generating process, resulting estimates will probably be less biased. This bias reduction cannot be verified and is not guaranteed, though—a flexible model that is overfitted or that incorporates instruments rather than confounders might fail to reduce bias, so caution is warranted.

Overfitting sounds bad. Do I need more data to fit a model that incorporates machine learning than I would need for a conventional regression model?

Holding all else equal, more flexible models need more data, and they need to be used cautiously to avoid overfitting. To account for this risk, machine-learning algorithms typically split samples when tuning parameter selection, effectively using less of the original data while training models. Nonetheless, more accurate model fit using machine-learning algorithms can decrease bias and increase precision, relative to misspecified parametric models (1). Moreover, some machine-learning algorithms, notably including techniques such as LASSO regression, perform well in settings with few observations. If you are concerned that you do not have enough data, it is wise to consult with machine-learning experts who might help you avoid algorithms that could more easily overfit your data and develop cross-validation plans that can make efficient use of the data you have.

Do I have to learn machine learning? Would all this effort benefit my work?

Nothing about using machine learning in causal inference changes the core value of epidemiologic thinking, which is (in our opinion) framing and formulating questions whose answers can improve public health (34, 35), with due consideration to systematic and random errors which may threaten the validity of the answers to those questions. The insights needed to formulate these questions rest more in conceptualizing population-level data-generating processes in order to design studies and interpret analytical results than they do in fancy estimation techniques (36). Moreover, while flexible model-fitting might help to better account for nonlinear confounding relationships as detailed above, some empirical simulations suggest that regression-based causal effect estimates are fairly robust to bias from statistical model misspecification (1).

The field of machine learning is actively developing, however, and the incremental benefits of flexible model-fitting may increase as the field matures. When the plausibility of outcome regression's linearity assumptions is a strong concern, machine learning provides a principled approach to model selection which allows us to more realistically capture uncertainty about model form and address scenarios with many potential confounders and limited background information in order to choose among them.

Could this all backfire, though? Could machine learning increase bias in my results?

It is unlikely that cautious use of appropriate machine-learning algorithms will increase bias. If you understand pitfalls of variable selection, are aware that the bias-variance tradeoff may not be for the parameter of interest, and can ensure that your sample is large enough for sample-splitting to reduce overfitting, you will probably avoid the worst-case scenarios.

However, machine learning is not just 1 tool, and different algorithms encode different assumptions about functional forms. It is possible that, for example, selecting a tree-based estimator to fit a smooth function could result in greater misspecification than would result from a typical regression model. In practice, we would expect that bias arising from an algorithm's failing to approximately represent the data-generating process will often, though not necessarily, result in artifacts that a careful analyst would notice. For example, inappropriate selection of an exposure instrument as a confounder might result in very large propensity scores (1), or overfitted predictive models might result in large differences between training set and test set cross-validation results, which could be caught as an analysis progresses. Another way to address model-choice concerns is through the use of ensemble models, which we discuss in the next question. Broadly, as in all epidemiology, cautious interpretation of results is warranted.

Will I get accurate confidence intervals?

Methods for estimating frequentist confidence intervals typically use math for populations that approach infinite sizes. This math, known as asymptotic theory, defines 2 terms that are key for estimating confidence intervals: consistency (a point estimate converges to the true value as sample size grows) and asymptotic normality (the standardized difference between point estimates and the true value converges to a normal distribution in very large samples) (37). For consistent and asymptotically normal estimators (e.g., the sample mean), asymptotic theory gives nice (and familiar!) estimates of variance and, hence, confidence intervals.

Informally, asymptotic normality and consistency both rely on the "information" in the data growing faster than the number of parameters. Many machine-learning algorithms prevent overfitting by penalizing parameters (optimizing the

Table 1. Common Categories of Machine-Learning Algorithms

Category	Examples	Benefits	Drawbacks
Generalized linear models	Linear regression, logistic regression	Familiarity among other researchers, extensive software support	Linearity assumptions, sensitivity to highly correlated variables
Penalized linear regression	LASSO, ridge regression, elastic net regression	Robustness within wide data sets	Linearity assumption
Tree-based methods	CART, random forest	Discontinuous effects, model flexibility	Risk of overfitting when discontinuity is not appropriate
Spline-based methods	LOESS smoothed regression, generalized additive models	Model flexibility	High data requirements, particularly when interactions are of interest

Abbreviations: CART, classification and regression trees; LASSO, least absolute shrinkage and selection operator; LOESS, locally weighted scatterplot smoothing.

bias/variance tradeoff; see "What is this bias/variance tradeoff thing?"), which introduces biases that do not diminish (or diminish too slowly to produce accurate confidence intervals (31)) in large samples. This in turn means that 95% confidence interval estimates may not cover the range that would include the true value in 95% of infinitely repeated samples. There is active research exploring how selected estimators may allow consistent and asymptotically normal estimators. However, this field is nascent, and no consensus has been reached about which algorithms will function well in actual data with small or moderate sample sizes (38, 39).

Given the lack of consensus, the safest option as of 2021 appears to be to use cross-fitting (31, 40), certain forms of doubly robust estimation (41, 42), or an approach embedding a higher-order influence function (38, 43). These options permit computation of confidence intervals supported by theory at the cost of more intricate programming.

Okay, okay, fine. How do I get started with machine learning?

There are numerous textbooks, online tutorials, coursework, and seminars available that provide background information on machine learning, and many of these touch on using machine learning in causal analyses. In particular, we would point readers to *An Introduction to Statistical Learning* (6) for background information on machine learning and to *Targeted Learning in Data Science* (44) for more details on incorporating machine learning into causal analyses.

There are myriad machine-learning algorithms; each has its own different benefits and drawbacks (6), and some empirical studies suggest that performance differences between some popular algorithms are minimal in scenarios approximating realistic epidemiologic data sets (1, 30, 45, 46). At a minimum, we would recommend familiarizing yourself with major classes of machine-learning methods and their benefits and drawbacks (Table 1).

Because some algorithms suit some problems much better than others, ensemble models that combine multiple algorithms are a common and defensible choice. Ensemble methods combine multiple machine-learning algorithms and are desirable in that they optimize prediction in a way that typically outperforms individual, component algorithms. In practice, ensemble methods trade computational time for flexibility. The best known ensemble methods at the moment are Super Learner (47–49), random forest (50), and XGBoost (51), which are flavors of the classes "stacking," "bootstrap aggregation," and "boosting," respectively (6).

Software for machine learning is improving quickly. As of May 2021, good general support for starting to work with machine learning is available in R (the *caret* package (52); R Foundation for Statistical Computing, Vienna, Austria), Python (the *scikit-learn* package (53); Python Software Foundation, Beaverton, Oregon), and SAS (46) (SAS Institute, Inc., Cary, North Carolina). These can help you if you want to compute propensity scores using machine learning and then conduct a conventional propensity-score–matched or inverse-probability–weighted analysis. Web Appendices 1–3 (available online at https://doi.org/10.1093/aje/kwab047) contain annotated sample code in R, Python, and SAS that uses classification and regression trees for computing propensity scores and inverse probability of treatment weights and then using the latter in analysis.

If you want an all-in-one package, TMLE features the most mature software for incorporating machine-learning approaches into causal inference. There are multiple packages for TMLE in R, but the *tmle* package is the simplest to start with (32). Support in SAS (46) and Stata (StataCorp LLC, College Station, Texas) is developing but less established. Web Appendix 4 contains annotated sample code in R using TMLE to estimate a causal effect. Code from all appendices is also available on GitHub (54).

CONCLUSIONS

Machine learning is a useful tool for epidemiologists and can be used not only to improve descriptive and predictive tasks such as measurement, imputation, and risk assessment but also in the service of causal effect estimation. Nonetheless, the core task of epidemiology—understanding

the abstract principles of inference from data and applying those principles to answer consequential questions about population health—is unchanged by the increasing incorporation of machine learning.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, School of Public Health, University of Washington, Seattle, Washington, United States (Stephen J. Mooney); Harborview Injury Research and Prevention Center, University of Washington, Seattle, Washington, United States (Stephen J. Mooney); and Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, North Carolina, United States (Alexander P. Keil, Daniel J. Westreich).

S.J.M. was supported by grant K99012868 from the National Library of Medicine. A.P.K. was supported by grant R01ES029531 from the National Institute of Environmental Health Sciences. D.J.W. was supported by grant DP2HD084070 from the Office of the Director of the National Institutes of Health and the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development.

S.J.M. thanks the University of Washington Statistical Learning Reading Group.

Conflict of interest: none declared.

REFERENCES

1. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med*. 2010;29(3):337–346.
2. 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(8):826–833.
3. Mooney SJ, Pejaver V. Big data in public health: terminology, machine learning, and privacy. *Annu Rev Public Health*. 2018;39:95–112.
4. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform*. 2007;11(2):59–77.
5. Friedman J, Hastie T, Tibshirani R. *The Elements of Statistical Learning*. New York, NY: Springer Publishing Company; 2001.
6. James G, Witten D, Hastie T, et al. *An Introduction to Statistical Learning*. New York, NY: Springer Publishing Company; 2013.
7. Darcy AM, Louie AK, Roberts LW. Machine learning and the profession of medicine. *JAMA*. 2016;315(6):551–552.
8. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA*. 2017;318(6):517–518.
9. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat Sci*. 2001;16(3):199–231.
10. Murphy KP. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press; 2012.
11. Robins JM. Data, design, and background knowledge in etiologic inference. *Epidemiology*. 2001;12(3):313–320.
12. van der Laan MJ, Rose S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, NY: Springer Science+Business Media; 2011.
13. Athey S, Imbens G. Machine learning methods for estimating heterogeneous causal effects [preprint]. *arXiv*. 2015. (doi: arXiv:1504.01132v3). Accessed May 7, 2021.
14. Díaz I. Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics*. 2020;21(2):353–358.
15. Nguyen QC, Sajjadi M, McCullough M, et al. Neighbourhood looking glass: 360^{o} automated characterisation of the built environment for neighbourhood effects research. *J Epidemiol Community Health*. 2018;72(3):260–266.
16. Naimi AI, Platt RW, Larkin JC. Machine learning for fetal growth prediction. *Epidemiology*. 2018;29(2):290–298.
17. Rittenhouse KJ, Vwalika B, Keil A, et al. Improving preterm newborn identification in low-resource settings with machine learning. *PLoS One*. 2019;14(2):e0198919.
18. Keil AP, Mooney SJ, Jonsson Funk M, et al. Resolving an apparent paradox in doubly robust estimators. *Am J Epidemiol*. 2018;187(4):891–892.
19. Pearl J. *Causality*. Cambridge, United Kingdom: Cambridge University Press; 2009.
20. Breskin A, Cole SR, Hudgens MG. A practical example demonstrating the utility of single-world intervention graphs. *Epidemiology*. 2018;29(3):e20–e21.
21. Hernán MA, Robins JM. *Causal Inference: What If.* Boca Raton, FL: Chapman & Hall/CRC Press; 2020.
22. Andrews B, Ramsey J, Cooper GF. Learning high-dimensional directed acyclic graphs with mixed data-types. *Proc Mach Learn Res*. 2019;104:4–21.
23. Alaa AM, van der Schaar M. Bayesian nonparametric causal inference: information rates and learning algorithms. *IEEE J Sel Top Signal Process*. 2018;12(5):1031–1046.
24. Petersen ML, van der Laan MJ. Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology*. 2014;25(3):418–426.
25. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B Methodol*. 1996;58(1):267–288.
26. Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163(12):1149–1156.
27. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar Behav Res*. 2011;46(3):399–424.
28. Pearl J. Invited commentary: understanding bias amplification. *Am J Epidemiol*. 2011;174(11):1223–1227.
29. Keil AP, Edwards JK. You are smarter than you think: (super) machine learning in context. *Eur J Epidemiol*. 2018;33(5):437–440.
30. Setoguchi S, Schneeweiss S, Brookhart MA, et al. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiol Drug Saf*. 2008;17(6):546–555.
31. Chernozhukov V, Chetverikov D, Demirer M, et al. Double/debiased machine learning for treatment and structural parameters. *Econometr J*. 2018;21(1):C1–C68.
32. Gruber S, van der Laan MJ. tmle: An R package for targeted maximum likelihood estimation. *J Stat Softw*. 2011;51(13):1–35.
33. Naimi AI, Mishler AE, Kennedy EH. Challenges in obtaining valid causal effect estimates with machine learning algorithms [preprint]. *arXiv*. 2020. (doi: arXiv:1711.07137v2). Accessed October 4, 2020.

34. Hernán MA, Taubman SL. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *Int J Obes (Lond)*. 2008;32(suppl 3):S8–S14.
35. Westreich D, Edwards JK, Rogawski ET, et al. Causal impact: epidemiological approaches for a public health of consequence. *Am J Public Health*. 2016;106(6):1011–1012.
36. Fox MP, Edwards JK, Platt R, et al. The critical importance of asking good questions: the role of epidemiology doctoral training programs. *Am J Epidemiol*. 2020;189(4): 261–264.
37. Newey WK, McFadden D. Large sample estimation and hypothesis testing. *Handb Econom*. 1994;4:2111–2245.
38. Robins JM, Li L, Mukherjee R, et al. Minimax estimation of a functional on a structured high-dimensional model. *Ann Stat*. 2017;45(5):1951–1987.
39. Cai W, van der Laan M. Nonparametric bootstrap inference for the targeted highly adaptive LASSO estimator [preprint]. *arXiv*. 2020. (doi: arXiv:1905.10299v2). Accessed October 4, 2020.
40. Bickel PJ. On adaptive estimation. *Ann Stat*. 1982;10: 647–671.
41. Benkeser D, Carone M, van der Laan M, et al. Doubly robust nonparametric inference on the average treatment effect. *Biometrika*. 2017;104(4):863–880.
42. van der Laan MJ. Targeted estimation of nuisance parameters to obtain valid statistical inference. *Int J Biostat*. 2014;10(1): 29–57.
43. Robins J, Li L, Tchetgen E, et al. Higher order influence functions and minimax estimation of nonlinear functionals. In: *Probability and Statistics: Essays in Honor of David A. Freedman*. Cleveland, OH: Institute of Mathematical Statistics; 2008:335–421.
44. van der Laan MJ, Rose S. *Targeted Learning in Data Science*. New York, NY: Springer Publishing Company; 2018.
45. Wainer J. Comparison of 14 different families of classification algorithms on 115 binary datasets [preprint]. *arXiv*. 2016. (doi: arXiv:1606.00930). Accessed October 4, 2020.
46. Keil AP, Westreich DJ, Edwards JK, et al. Super learning in the SAS system [preprint]. *arXiv*. 2019. (doi: arXiv:1805.08058v3). Accessed October 4, 2020.
47. Polley EC, van der Laan MJ. *Super Learner in Prediction*. (U.C. Berkeley Division of Biostatistics Working Paper Series, working paper 266). Berkeley, CA: Division of Biostatistics, University of California, Berkeley; 2010. http://biostats.bepress.com/ucbbiostat/paper266/. Accessed October 4, 2020.
48. Naimi AI, Balzer LB. Stacked generalization: an introduction to super learning. *Eur J Epidemiol*. 2018;33(5):459–464.
49. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol*. 2007;6(1):e25.
50. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
51. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: Association for Computing Machinery; 2016:785–794.
52. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw*. 2008;28(5):1–26.
53. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12(85): 2825–2830.
54. Keil AP. CIRL-UNC/12MLquestions. https://github.com/CIRL-UNC/12MLquestions. Published July 12, 2020. Accessed May 7, 2021.