American Journal of Epidemiology

Vol. 00, No. 00
DOI: 10.1093/aje/kwab048
Advance Access publication:

Invited Commentary

Invited Commentary: Machine Learning in Causal Inference—How Do I Love Thee? Let Me Count the Ways

Laura B. Balzer* **and Maya L. Petersen**

* Correspondence to Dr. Laura B. Balzer, Department of Biostatistics and Epidemiology, School of Public Health and Health Sciences, University of Massachusetts Amherst, 427 Arnold House, Amherst, MA 01003 (e-mail: lbalzer@umass.edu).

Initially submitted November 30, 2020; accepted for publication February 4, 2021.

In this issue of the *Journal*, Mooney et al. (*Am J Epidemiol*. 2021;000(00):000–000) discuss machine learning as a tool for causal research in the style of Internet headlines. Here we comment by adapting famous literary quotations, including the one in our title (from "Sonnet 43" by Elizabeth Barrett Browning (*Sonnets From the Portuguese*, Adelaide Hanscom Leeson, 1850)). We emphasize that any use of machine learning to answer causal questions must be founded on a formal framework for both causal and statistical inference. We illustrate the pitfalls that can occur without such a foundation. We conclude with some practical recommendations for integrating machine learning into causal analyses in a principled way and highlight important areas of ongoing work.

Common problem across all of statistics: Scientists running multiple unclear models until the one that individual wants turns up.

Instead, the model should be clearly defined and planned out before running to ensure principled learning from the data.

ML that is used with causal inference, however, should be based in guiding principles of the data and formal causal frameworks, and are themselves not to be used with causal inference without much thought and planning.

However, although this paper guides people to think about their work (harm reduction), does it encourage them to think enough? Instead, should the knowledge one should have to properly run causal inference studies be discussed rather than just encourage slight more (and still not enough) thought?

Editor's note: *The opinions expressed in this article are those of the authors and do not necessarily reflect the views of the* American Journal of Epidemiology.

To [ML] or not to [ML]—that is [not] the question

***—William Shakespeare* (1)**

Machine learning (ML) has become ubiquitous in public health and epidemiologic research (2, 3). Supervised learning algorithms, which estimate the expected value of an observed variable given a set of other measured variables, are commonly used to improve predictions (4–6). In epidemiology, however, we often ask causal questions—questions about what outcomes would look like under alternative hypothetical conditions (e.g., a change in how a treatment was assigned or an exposure distributed) (7). As Mooney et al. (8) discuss in their accompanying article, supervised ML also offers the promise of better answers to these causal questions.

The need for ML is clear, particularly in modern data ecosystems where we often face dozens of, if not more, confounding variables. Stratification-based approaches are typically ill-defined because of empty or sparse cells; we rarely have the knowledge to specify a correct parametric regression a priori, and data-snooping (fitting a series of estimators and selecting the "best" in an ad hoc manner) or *P*-hacking (conscious or not) undermines the foundations of statistical inference. In place of these unsatisfactory alternatives, ML offers a principled and prespecified way to flexibly learn from the data.

While ML is often an essential ingredient for causal inference, even the best ML algorithm may yield wildly misleading answers to causal questions if the rest of the recipe is ignored. We cannot simply accessorize our ML-based predictions with causal assumptions (e.g., no unmeasured confounding) or statistical concepts (e.g., a bootstrap) after the fact. Instead, ML algorithms must be carefully integrated within a formal framework for causal and statistical inference.

When I'd heard the learn'd [epidemiologist]

***—Walt Whitman* (9)**

Researchers are sometimes worried that ML will supplant human expertise and experience in causal inference research

(10). In contrast, such knowledge forms the essential foundation for using ML to answer causal questions. Consider, for example, the following steps of the Causal Roadmap, one of several frameworks for causal and statistical inference (11–16).

1. State the research question, including the target population, primary exposure, primary outcome, and scale of comparison.
2. Specify a causal model, such as a directed acyclic graph (17), to represent causal relationships between key variables, including potential sources of bias (e.g., confounding, selection, missing data, censoring).
3. Translate the research question into a well-defined causal parameter, a summary measure of the distribution of the counterfactual (potential) outcomes (e.g., the difference in participants' expected outcomes under both exposed and unexposed conditions).
4. Specify what data are available in actuality and the link between the causal and statistical models.
5. Identify: Translate the causal parameter to a statistical parameter—a summary measure of the observed data distribution—by critically evaluating the assumptions encoded in the causal model (together with a statistical assumption of adequate data support).
6. Estimate: Obtain point estimates and inference for the corresponding statistical parameter (e.g., with matching, G-computation, inverse weighting, augmented inverse weighting, or targeted maximum likelihood estimation (TMLE) with Super Learner).
7. Interpret results. Causal interpretation is warranted only when identifiability assumptions hold (step 5).

ML doesn't replace knowledge and understanding of the data and underlying problem. This has to be established to even BEGIN to apply ML to a problem

ML, and more generally statistical estimation, only plays a role in step 6. (Following Mooney et al. (8), we do not address causal discovery algorithms in this commentary.) The other steps of the Causal Roadmap rely almost exclusively on human knowledge and expertise.

To illustrate, consider the directed acyclic graph and corresponding nonparametric structural equations in Figure 1. Here, $W = \{W1, W2, W3\}$ represents measured confounders, I an instrumental variable, M another preexposure variable but not a confounder, A the exposure indicator, Y the outcome, $U1$ unmeasured common causes of M and A, and $U2$ unmeasured common causes of M and Y. Given the observed data $O = (W, I, M, A, Y)$, a naive approach to evaluate the causal effect of A on Y might start by applying a sophisticated algorithm to estimate the conditional expectation of the outcome Y given its measured past (W, I, M, A), or alternatively to estimate the conditional probability of the exposure A given its measured past (W, I, M), that is, the propensity score. While such an approach is subject to a number of possible pitfalls, ignoring the causal model is arguably the most critical, because we end up estimating a statistical parameter that differs meaningfully from any interpretable causal effect, even when the "no unmeasured confounding" assumption holds, as in Figure 1. As the simulation study presented in Table 1 illustrates, this approach can yield deeply misleading inferences for the causal effect of interest.

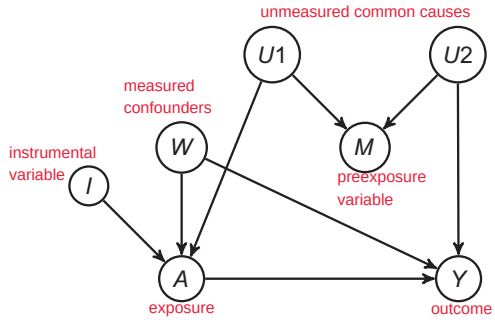

Figure 1. Example of a directed acyclic graph in which $W = \{W1, W2, W3\}$ represents measured confounders, I an instrumental variable, M another preexposure variable but not a confounder, A the exposure indicator, Y the outcome, $U1$ unmeasured common causes of M and A, and $U2$ unmeasured common causes of M and Y. The unmeasured causes of the confounders U_W and of the instrumental variable U_I are independent of the others and are not shown. The corresponding nonparametric structural equations are $W = f_W(U_W)$; $I = f_I(U_I)$; $M = f_M(U_M)$; $A = f_A(W, I, U_A)$; and $Y = f_Y(W, A, U_Y)$.

Mooney et al. refer to these errors as "causal model misspecification" (8). While incorrectly specified causal models can certainly lead to identification errors, perhaps a more common error is "causal model neglect": the failure to use causal knowledge to carefully specify a target statistical parameter before proceeding to estimation. If, instead, we had followed the Causal Roadmap, the identification step would have led us to the following statistical parameter, expressed as the G-computation formula

$$\mathbb{E}\left[\mathbb{E}\left(Y|A=1,W\right)\right] - \mathbb{E}\left[\mathbb{E}\left(Y|A=0,W\right)\right] \quad (1)$$

or, equivalently in inverse-weighted form,

$$\mathbb{E}\left[\frac{\mathbb{I}\left(A=1\right)}{\mathbb{P}\left(A=1|W\right)}Y\right] - \mathbb{E}\left[\frac{\mathbb{I}\left(A=0\right)}{\mathbb{P}\left(A=0|W\right)}Y\right]. \quad (2)$$

In this setting, the instrumental variable I and the M-biasing variable M can and should be ignored for purposes of estimation and inference (17–20). As Table 1 shows, both G-computation and inverse weighting exhibited minimal bias and good confidence interval coverage when their adjustment sets followed from the identification result (step 5) and when a correctly specified parametric regression was used in estimation (step 6). Of course, in practice, our knowledge is generally insufficient to enable correct specification of parametric regressions, and ML is needed to help address this challenge.

While Table 1 may seem like an extreme example, errors of "causal model neglect" commonly occur when heeding advice to adjust for all preexposure variables or when the exposure is time-varying (i.e., longitudinal) (21, 22). It may seem obvious, but it is nonetheless worth stating explicitly: Background knowledge remains *the* foundation of causal identification (step 5), and ML *cannot* uncover cause-and-effect if this foundation is weak. Instead, ML provides an essential tool in the design of statistical estimators able to

When information is used to define the model to build, bias is reduced

Table 1. Results From a Simulation Study Illustrating the Consequences of Neglecting the Causal Model[a]

Estimator	Mean, %	Bias, %	Coverage, %	Relevant Regression[b]
Unadjusted	15.6	1.3	84.6	$Y \sim A$
Naive implementation				
G-computation	17.9	3.5	75.0	$Y \sim W1 + W2 + W3 + I + M + A$
Inverse weighting	25.8	11.5	81.2	$A \sim W1 + W2 + W3 + I + M$
Roadmap-informed				
G-computation	14.4	0.1	96.4	$Y \sim W1 + W2 + W3 + A$
Inverse weighting	14.3	0.0	100.0	$A \sim W1 + W2 + W3$

[a] We consider the average treatment effect—defined as the difference in the expected counterfactual outcome under the exposure $\mathbb{E}(Y_1)$ and under no exposure $\mathbb{E}(Y_0)$, and equal to 14.3% in this simulation (38). Over 1,000 repetitions of the data-generating process, which is compatible with Figure 1, we show the mean point estimate, bias (average deviation between point estimate and true effect), and coverage (proportion of times the calculated 95% confidence interval contained the true effect) for the following estimators: unadjusted, with G-computation naively implemented (regressing the outcome on the measured past); inverse weighting naively implemented (regressing for exposure on the measured past); G-computation informed by the Causal Roadmap (regressing the outcome on exposure and confounders); and inverse weighting informed by the Roadmap (regressing the exposure on the confounders).

[b] ($W1$, $W2$, $W3$) are confounders, I is an instrumental variable, M is a preexposure variable but not a confounder, A is the exposure indicator, and Y is the outcome.

provide valid inferences when faced with realistic statistical models: models that accurately reflect our limited knowledge.

What's in a [statistical model]?

***—William Shakespeare* (23)**

Mooney et al. refer to errors stemming from reliance on parametric assumptions as "statistical model misspecification" (8). As before, a more apt term might be "statistical model neglect": failure to respect our statistical knowledge during the estimation process. Such errors can be avoided by ensuring that the statistical model, formally defined as the set of all possible distributions of the observed data, only represents real knowledge—*not* assumptions made for convenience at the estimation stage (24). Step 4 of the Causal Roadmap guarantees that our knowledge of the data-generating process (as opposed to wished-for simplifications) is carried through to the statistical model. For example, we often assume that the observed data are generating by sampling N times from a data-generating process compatible with the causal model (24). Under this assumption, the causal model implies the statistical model, characterizing the set of possible distributions of the observed data. In practice, few or no restrictions are placed on this set, yielding a semiparametric or nonparametric statistical model, respectively.

For example, the causal model in Figure 1 does not encode parametric knowledge about the functional forms of the relationships between variables. Focusing on the propensity score, for example, the causal model encodes our limited knowledge that the exposure A is some unknown function of the confounders W, the instrument I, and unmeasured factors U_A. The causal model does not state, for example, that the conditional probability of the exposure A is accurately described by a main-terms logistic function of the confounders W and instrument I. Instead, such a main-terms function is just one of many possible ways that the exposure A could be generated from (W, I, U_A). During statistical estimation (Roadmap step 6), using a statistical model that reflects this uncertainty provides the foundation for accurate inferences.

Not all those who wander are lost

***—J.R.R. Tolkien* (25)**

Here is where the power and necessity of ML-based approaches become clear. Respecting the limits of our knowledge forces us to confront very large statistical models—for example, those without functional-form restrictions on the conditional probability of exposure given confounders or on the expected outcome given exposure and confounders. In doing so, we are empowered to dismiss George Box's quotation, "All models are wrong" (26, p. 792).

Instead, we can joyfully proclaim, "My statistical model correctly describes reality." However, in doing so, we also face new challenges for statistical estimation and inference. In particular, we are forced to leave behind the familiar comforts of parametric regressions and strike out on a journey through the vast space of distributions contained in our statistical model. Respect for our statistical model means that we can, and indeed must, explore a wide range of relationships between exposure and confounders as well as relationships between the outcome, exposure, and confounders.

Supervised ML provides the means to conduct this exploration in a powerful, principled, and fully prespecified

More emphasis is needed on integrating expert knowledge - this is one of the most important requirements for using ML

manner. Ensemble methods, such as Super Learner (27, 28), are particularly promising, because they use K-fold cross-validation (i.e., sample-splitting) to build the optimal weighted combination of predictions from a set of candidate algorithms. Importantly, background knowledge can again play a key role through the inclusion of expert-guided interaction terms or other features in the predictor set and parametric regressions in the algorithm set.

Do you know anything on earth which has not a dangerous side if it is mishandled and exaggerated?

***—Sir Arthur Conan Doyle* (29)**

decrease and bias and variance from ML, but the decrease in variance will be much slower than decrease in bias. thus as bias decreases the variance does so much slower.

Nonetheless, even when appropriately confined to the estimation stage of the Causal Roadmap (step 6), ML is not without dangers. First and foremost, there may be a temptation to use ML-based predictions in singly robust methods, such as G-computation or inverse weighting. This approach may well outperform singly robust methods relying on misspecified parametric regressions and, assuming that the ML algorithms are flexible enough, provides the benefit of decreasing bias as sample size increases. The decrease in bias, however, is typically too slow to offset the corresponding decrease in variance, resulting in the potential for misleading statistical inference (i.e., lower than nominal confidence interval coverage).

One way to understand this challenge is that ML-based predictions are generated on the basis of minimizing some loss function corresponding to the supervised learning task. For example, ML may be used to do the best possible job predicting the outcome for all possible values and combinations of the exposure and confounders, while the true value of the G-computation formula (equation 1) is just one number (equal to the average treatment effect under the identifiability assumptions). In other words, a full prediction function is a different estimation goal than the G-computation formula and thereby has a different optimal bias-variance tradeoff. Equally important, there is no theory to support that the central limit theorem applies to the resulting estimators. Therefore, the 95% confidence intervals resulting when using ML with G-computation or inverse weighting should be regarded with suspicion.

TMLE is becoming much more apparent and used in the literature as of late. as a framework it is very easy to use - perhaps encouraging use when not fully understood?

These challenges have inspired the development of several double-robust estimators, such as augmented inverse weighting and TMLE (24, 30–34). These approaches can combine ML-based estimates of the expected outcome and the propensity score to achieve a number of desirable asymptotic properties, including the construction of valid 95% confidence intervals under regularity conditions. Double robust estimators employing sample-splitting, such as cross-validated TMLE, can help to ensure that the conditions required for valid statistical inference are met in practice (34–36).

An additional practical challenge is selection and implementation of the ML algorithm best suited for the current problem. Approaches like Super Learner allow us to formally explore a variety of algorithms (including the same algorithm with different tuning parameters). However, the performance of an ensemble approach is driven by the set of algorithms considered. We recommend incorporating expert knowledge and including simple parametric regressions, together with more flexible approaches. With hierarchical or repeated-measures data, it is essential to sample-split on the independent unit (e.g., the individual in longitudinal settings). Finally, if the dependent variable is rare, we recommend stratifying on the outcome before sample-splitting to maintain roughly the same prevalence in each split. Of course, these recommendations are just that—recommendations. Implementation of any ML algorithm with real data always raises complex challenges, and careful examination of the default settings of any statistical computing package (as well as, ideally, performance evaluation using simulation) is warranted.

I [thoughtfully ML], therefore I am

***—René Descartes* (37)**

In summary, recent advances in ML provide a tremendous opportunity to improve epidemiologic research by reducing or (better yet) eliminating our reliance on unrealistically restrictive statistical assumptions. However, this opportunity is only afforded when our analyses are guided by epidemiologic principles, formal causal frameworks, and statistical theory (38).

ACKNOWLEDGMENTS

Author affiliations: Department of Biostatistics and Epidemiology, School of Public Health and Health Sciences, University of Massachusetts Amherst, Amherst, Massachusetts, United States (Laura B. Balzer); and Division of Biostatistics, School of Public Health, University of California, Berkeley, Berkeley, California, United States (Maya L. Petersen).

L.B.B. thanks Thomas Hungerford and Bruce Coffin (Westover School, Middlebury, Connecticut) for inspiring the love of literature and learning in countless generations of students.

Computer code for reproducing the study results is available on GitHub (39).

Conflict of interest: none declared.

REFERENCES

1. Shakespeare W. *Hamlet*, *First Folio*. London, United Kingdom: Stationers Company; 1623.
2. Mooney SJ, Pejaver V. Big data in public health: terminology, machine learning, and privacy. *Annu Rev Public Health*. 2018;39:95–112.
3. Bi Q, Goodman KE, Kaminsky J, et al. What is machine learning? A primer for the epidemiologist. *Am J Epidemiol*. 2019;188(12):2222–2239.
4. Rose S. Mortality risk score prediction in an elderly population using machine learning. *Am J Epidemiol*. 2013; 177(5):443–452.
5. Baćak V, Kennedy EH. Principled machine learning using the super learner: an application to predicting prison violence. *Sociol Methods Res*. 2019;48(3):698–721.

6. Marcus JL, Sewell WC, Balzer LB, et al. Artificial intelligence and machine learning for HIV prevention: emerging approaches to ending the epidemic. *Curr HIV/AIDS Rep*. 2020;17(3):171–179.
7. Pearl J. Causal inference in statistics: an overview. *Statist Surv*. 2009;3:96–146.
8. Mooney SJ, Keil AP, Westreich DJ, et al. Thirteen questions about using machine learning in causal research (you won't believe the answer to number 10!). *Am J Epidemiol*. 2021; 000(00):000–000.
9. Whitman W. *Drum-Taps*. New York, NY: Peter Eckler; 1865.
10. Keil AP, Edwards JK. You are smarter than you think: (super) machine learning in context. *Eur J Epidemiol*. 2018;33(5): 437–440.
11. Petersen ML, van der Laan MJ. Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology*. 2014;25(3):418–426.
12. Petersen ML, Balzer LB. Introduction to causal inference. www.ucbbiostat.com. Published August 2014. Updated December 2018. Accessed February 1, 2021.
13. Petersen ML. Commentary: applying a causal road map in settings with time-dependent confounding. *Epidemiology*. 2014;25(6):898–901.
14. Balzer L, Petersen M, van der Laan MJ. Tutorial for causal inference. In: Buhlmann P, Drineas P, Kane M, et al., eds. *Handbook of Big Data*. Boca Raton, FL: Chapman & Hall/CRC Press; 2016:361–386.
15. Tran L, Yiannoutsos CT, Musick BS, et al. Evaluating the impact of a HIV low-risk express care task-shifting program: a case study of the targeted learning roadmap. *Epidemiol Methods*. 2016;5(1):69–91.
16. Saddiki H, Balzer LB. A primer on causality in data science. *J Société FrançStatist*. 2020;161(1):67–90.
17. Pearl J. *Causality: Models, Reasoning and Inference*. 2nd ed. New York, NY: Cambridge University Press; 2009.
18. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*. 2003;14(3):300–306.
19. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15(5): 615–625.
20. Liu W, Brookhart MA, Schneeweiss S, et al. Implications of M bias in epidemiologic studies: a simulation study. *Am J Epidemiol*. 2012;176(10):938–948.
21. Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy worker survivor effect. *Math Model*. 1986;7:1393–1512.
22. Robins JM, Hernán MA. Estimation of the causal effects of time-varying exposures. In: Fitzmaurice G, Davidian M, Verbeke G, et al., eds. *Longitudinal Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC; 2009:553–597.
23. Shakespeare W. *Romeo and Juliet*, *First Folio*. London, United Kingdom: Stationers Company; 1623.
24. van der Laan MJ, Rose S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, NY: Springer Publishing Company; 2011.
25. Tolkien JRR. *The Fellowship of the Ring*. London, United Kingdom: George Allen & Unwin; 1954.
26. Box GEP, ed. Science and statistics. *J Am Stat Assoc*. 1976; 71(356):791–799.
27. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol*. 2007;6:Article25.
28. Naimi AI, Balzer LB. Stacked generalization: an introduction to super learning. *Eur J Epidemiol*. 2018;33(5):459–464.
29. Doyle AC. *The Land of Mist*. London, United Kingdom: Hutchinson & Co. (Publishers) Ltd.; 1926.
30. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. 1994;89(427):846–866.
31. Robins JM. Robust estimation in sequentially ignorable missing data and causal inference models. In: *1999 Proceedings of the American Statistical Association*. Alexandria, VA: American Statistical Association; 2000: 6–10.
32. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics*. 2005;61: 962–972.
33. van der Laan MJ, Rose S. *Targeted Learning in Data Science*. New York, NY: Springer Publishing Company; 2018.
34. 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.
35. Zheng W, van der Laan MJ. Cross-validated targeted minimum-loss-based estimation. In: van der Laan MJ, Rose S, eds. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, NY: Springer Publishing Company; 2011:459–474.
36. Benkeser D, Carone M, van der Laan MJ, et al. Doubly robust nonparametric inference on the average treatment effect. *Biometrika*. 2017;104(4):863–880.
37. Descartes R. *Discours de la Méthode pour Bien Conduire sa Raison, et Chercher la Vérité dans les Sciences*. Leiden, the Netherlands: Johannes Maire; 1637.
38. 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.
39. Balzer L. MachineLearningLove. https://github.com/LauraBalzer/MachineLearningLove. Published January 24, 2021. Accessed February 2, 2021.