Statistical Science
2020, Vol. 35, No. 3, 499–502
https://doi.org/10.1214/20-STS773
Main article: https://doi.org/10.1214/19-STS735

Comment: Automated Analyses: Because We Can, Does It Mean We Should?

Susan M. Shortreed and Erica E. M. Moodie

1. INTRODUCTORY REMARKS

We commend Benkeser, Cai and van der Laan (2020) for their interesting proposal and efforts to further automate the machinery of collaborative targeted minimum loss estimation (TMLE). Reducing human impact on an analysis, that is, to circumvent the need for analysts to "select an increasingly complex sequence of estimators [...] and implement each of these" is an important goal that could bring us closer to reproducible and transparent research. We agree that striving for estimators which have stable properties is a benefit, and practical positivity violations can render many estimators "erratic" or "nonrobust." In the examples, the authors showcase success in constructing data-specific robust estimators with well-behaved properties.

Petersen et al. (2012) describe TMLE as "an explicit trade-off [that is ideally] made in a systematic way rather than on an ad hoc basis at the discretion of the investigator." No statistical or machine learning-based approach is exempt from human-made decisions. For example, in ensemble-based machine learning methods, often used in conjunction with TMLE, the analyst must choose which methods to include in the ensemble learner, select hyperparameter values (e.g., random forests minimum node size), and select the number of folds for cross-validation. The question then arises as to whether it does, or should, trouble the scientific community that TMLE is less automated than we might think.

Here, we wish to probe two fundamental questions: Should automation and data-driven analyses be preferred when inferential, rather than predictive, analyses are undertaken? For example, is a data-adaptive *estimand* or an a priori human-defined estimand preferred? What do we lose by automating an increasing number of steps of scientific discovery?

Susan M. Shortreed is Senior Investigator Kaiser Permanente Washington Heath Research Institute, 1730 Minor Ave, Suite 1300, Seattle, Washington 98101, USA (e-mail: susan.m.shortreed@kp.org). Erica E. M. Moodie is William Dawson Scholar and Associate Professor of Biostatistics, McGill University, 1020 Pine Ave W., Montreal, Quebec, Canada H3A 1A2 (e-mail: erica.moodie@mcgill.ca).

2. REDUCING THE IMPACT OF HUMAN DECISIONS ON SCIENCE: AUTOMATION IN CAUSAL INFERENCE

Recent method developments mean statisticians have several tools to reduce confounding bias in treatment effects estimated from observational or nonexperimental data. However, this can result in an analyst at the computer choosing which variables to include as confounders, which approach to use to account for confounding, and which participants to include in the analytic data set. This scenario is ripe for honest mistakes and, in the worst case, can result in data dredging to find any "statistically significant" (i.e., publishable) results.

Human fallibility is not a new concept, though its role in statistical analyses has only recently been fully appreciated (Veldkamp, 2017). Focus on replicability and reproducibility in science (Peng, 2015, Baker, 2016) has led to improvements in documentation and open access sharing, particularly in the statistical sciences, where many journals insist that manuscripts are accompanied by code implementing the method or analysis. Documentation and sharing of data and code does not remove human-decisions from analyses, but it does, hopefully, reduce the negative impact human-made decisions can have on science through transparency.

In inferential statistics, various approaches have been proposed to minimize the impact of human decisions on study results. A long-standing approach, common in randomized trials, is prespecifying scientific questions and analytic plans. There is a growing movement to publish planned analyses for all inferential studies, including observational studies (Williams et al., 2010, Loder, Groves and Macauley, 2010, Lancet Editors, 2010, Hernán and Robins, 2016), but this is hardly ubiquitous. An alternative approach is to conduct as much of the analysis as possible blinded to study outcomes, making it difficult to skew study results with analytic decisions made along the way (Rubin, 2001, 2008). This approach has strengths, but has been shown to have reduced statistical efficiency compared to approaches that utilize outcome information in the entire analytic process (Greenland, 2008, Rotnitzky, Li and Li, 2010, Shortreed and Ertefaie, 2017, Ju, Benkeser and van der Laan, 2019). Recently, Schuemie et al. (2018) proposed a new paradigm for analyzing large clinical databases that analyzes multiple questions at once and requires the inclusion of "negative controls" (i.e., effects widely believed to be null) so

the operating characteristics of the analytic approach are better characterized.

Automation, common in machine learning, is another approach to removing human bias from an analysis. Machine learning tools were originally developed, and continue to be attractive, for their tremendous predictive abilities. Early in the development of these approaches, overfitting to the training data was recognized as a concern. Standard practice evolved such that tuning parameter selection is now made through cross-validation or similar approaches using an objective measure. It is also standard to use an independent validation or test set to accurately quantify the algorithm's performance. If only one data set is available for training and estimating model performance, approaches exist to address optimism in estimated model performance (Efron, 1983, Efron and Tibshirani, 1993, Smith et al., 2014).

TMLE is a method for estimating and making inference about treatment effects. This heavy statistical machinery targets a specific parameter of interest and leverages data to estimate this parameter efficiently. Applications of TMLE often make use of machine learning methods, specifically ensemble approaches, for flexible estimation of necessary models. Super learning was initially used for both outcome and treatment models in TMLE; only recently has it been discovered that exceptional prediction in the treatment model is not only unnecessary but potentially harmful (Alam, Moodie and Stephens, 2019, Pirracchio and Carone, 2018). This is not true for the outcome model, where flexibility and minimal-error prediction is desirable, underscoring that the goals of prediction and causal inference can differ, and may require different tools.

Biased treatment effect estimation, often in the direction of a type I error, is to inference, what unrealistic optimism of model performance is to prediction. Statistical efficiency can be improved by accounting for chance imbalances that occur in a particular study (Rotnitzky, Li and Li, 2010); yet allowing scientists to iteratively alter analyses after looking at study results changes the (already potentially problematic) meaning of p-values and can lead to publication of chance findings rather than generalizable results. Statisticians have long balanced the tension between bias and variance; we must also balance the tension between protecting against type 1 errors and improving statistical efficiency/robustness.

3. MOVING THE GOALPOSTS, OR JUST BEING PRACTICAL?

Human input has traditionally driven the definition of the population of interest, which in turn defined the analytic sample. The positivity, or experimental treatment assignment, assumption is central to estimating treatment effects from observational data (Rosenbaum and Rubin, 1983, Little and Rubin, 2000). Petersen et al. (2012) put a focused lens on the positivity assumption, and proposed a straightforward bootstrap-based diagnostic to assess it. The authors also laid out several approaches to dealing with positivity violations. One such approach was to change the variables included in the confounder set, thereby possibly trading bias for a reduction in the violation of the assumption. Another approach was to exclude from the sample individuals who, based on covariate values, might "always" receive only one of the treatment choices. Lastly, Petersen et al. (2012) propose choosing from among a family of estimands based on empirical evaluation of bias, noting that "researchers may be happy to settle for a better estimate of a less interesting parameter"—a data-adaptive approach that is similar in principle to that of Benkeser, Cai and van der Laan (2020). We support careful and introspective thought from the statistical community on how to best improve the reproducibility and reliability of science, and avoid erratic or unstable estimates. Nevertheless, questions remain about whether a wholesale move toward automation may undermine the process of scientific inquiry by shifting not only the answering of questions to the machine, but also the asking of them.

As statisticians, we wonder how to assess the replicability of scientific findings based on data-adaptive choices in the parameter of interest. It is unclear which is more valid: an unstable estimate of an estimand that is the same across samples or a more stable estimate for an estimand that changes across samples. We appreciate the desire to automate procedures to reduce human impact, but should we allow the data to drive the scientific question, that is, the choice of estimand, rather than the converse? Are we moving the goalposts and choosing to provide an estimate we can using the data we have, rather than seeking novel methods or different data to answer the question we want?

4. HUMAN MISTAKES, MACHINE MISTAKES; NEITHER ARE INFALLIBLE BOTH ARE USEFUL

While humans are indeed fallible and biased, computer-based algorithms have also made their fair share of "mistakes." The now infamous failure of Google Flu Trends to predict the 2009 H1N1 pandemic, followed by the gross overprediction of patient visits for flu-like symptoms in 2013, has called into question the value of automated algorithms over more traditional approaches that better account for the data structure (Lazer et al., 2014). The Apollo landing is another well-known example of where human intervention, working to counteract automation, saved the day: Armstrong and Aldrin had to find and manually land at an alternate landing spot when the lunar module was pushed off course (O'Neill).

Machine learning algorithms can also *propagate* disparities, where the original hope was they would reduce or remove said biases. A recent simulation study highlighted predictive policing using models that did not include race can still result in policing strategies that disproportionately affect nonwhites (Lum and Isaac, 2016). This arose because the algorithms were trained on historical data affected by existing biases in policing patterns. Biases in machine learning algorithms have been demonstrated in a variety of settings, including health care (Obermeyer et al., 2019). Especially when working with big data resources, data anomalies can lead to erroneous study results. For example, a study team examined health plan disenrollment (i.e., losing or changing health insurance) and risk of suicide death. They observed a strong relationship even after controlling for a number of confounders. After some data sleuthing, it was discovered that in this database, time of health plan disenrollment was often retrospectively assigned to the beginning of the month that a death occurred; thus, the data were not suited to estimate this relationship (G Simon personal communication).

Errors in automation often stem from a lack of full understanding and encoding of the entire data generating process when estimating the procedure. Human input is essential to ensure both methods and data are used appropriately to address not only confounding but other potential biases arising from the processes of data collection and measurement, including nonrandom selection of the study sample, measurement error and covariate-informed measurements, visits and drop-out. Automating analyses can produce reliable results; it can also lead us to overlook important details of the analytic sample. It is important to ask how automation should best be used to create reproducible and replicable research and to consider how both the analysis and the data on which it relies (data generating process) impact study results.

5. CONCLUDING THOUGHTS

Advances in computing power and predictive ability have transformed what is possible in statistical analyses of both big and small data. Building on these strengths to automate analyses to ensure reproducibility is a noble goal. Reducing the impact of human-bias is paramount to good science. However, not all of science (or analysis) lends itself to automation, nor does automation always circumvent human-bias. The question of what we should automate and why is as important as how it should be done (i.e., the methods selected). We applaud Benkeser, Cai and van der Laan (2020) for proposing an interesting method that offers robustness and stability, two important statistical properties. We would like to also encourage the authors and the wider scientific community to discuss the question: When we should automate? Human impact in any real-world data analysis is unavoidable; sometimes this human impact can prevent an automated algorithm from going astray and sometimes automation can stop humans from impacting results in way that compromises scientific integrity.

ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Institute of Mental Health of the National Institutes of Health under Award Number R01 MH114873. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. EEMM acknowledges a chercheur boursier senior career award from the Fonds de recherche du Québec–Santé.

The authors are supported by NIMH R01 MH114873.

REFERENCES

ALAM, S., MOODIE, E. E. M. and STEPHENS, D. A. (2019). Should a propensity score model be super? The utility of ensemble procedures for causal adjustment. *Stat. Med.* **38** 1690–1702. MR3934814 https://doi.org/10.1002/sim.8075

BAKER, M. (2016). Is there a reproducibility crisis? A nature survey lifts the lid on how researchers view the 'crisis rocking science and what they think will help. *Nature* **533** 452–454.

BENKESER, D., CAI, W. and VAN DER LAAN, M. (2020). A nonparametric super-efficient estimator of the average treatment effect. *Statist. Sci.* **35** 484–495.

EFRON, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Amer. Statist. Assoc.* **78** 316–331. MR0711106

EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap*. *Monographs on Statistics and Applied Probability* **57**. CRC Press, New York. MR1270903 https://doi.org/10.1007/978-1-4899-4541-9

GREENLAND, S. (2008). Invited commentary: Variable selection versus shrinkage in the control of multiple confounders. *Am. J. Epidemiol.* **167** 523–9; discussion 530–1. https://doi.org/kwm355[pii] 10.1093/aje/kwm355

HERNÁN, M. A. and ROBINS, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *Am. J. Epidemiol.* **183** 758–764. https://doi.org/10.1093/aje/kwv254

JU, C., BENKESER, D. and VAN DER LAAN, M. J. (2019). Robust inference on the average treatment effect using the outcome highly adaptive lasso. *Biometrics*. https://doi.org/10.1111/biom.13121

LANCET EDITORS (2010). Should protocols for observational studies be registered? *Lancet* **375** 348.

LAZER, D., KENNEDY, R., KING, G. and VESPIGNANI, A. (2014). Big data. The parable of Google flu: Traps in big data analysis. *Science* **343** 1203–1205. https://doi.org/10.1126/science.1248506

LITTLE, R. J. and RUBIN, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches. *Annu. Rev. Public Health* **21** 121–145. https://doi.org/10.1146/annurev.publhealth.21.1.121

LODER, E., GROVES, T. and MACAULEY, D. (2010). Registration of observational studies: The next step toward research transparency. *BMJ* **340** c950. https://doi.org/10.1136/bmj.c950

LUM, K. and ISAAC, W. (2016). To predict amd serve? *Significance*.

O'NEILL, I. Apollo Landing Terrifying Moments. https://urldefense.com/v3/__https://www.history.com/news/apollo-11-moon-landing-terrifying-moments__;!7TrXCGkIugIq!4-s2OGm68zL4bh2pdYQHkMN56a89SJUDHoYeoBUWwSbZCd8WnK_Icg4Orhxehkk87oE.

OBERMEYER, Z., POWERS, B., VOGELI, C. and MULLAINATHAN, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366** 447–53.

PENG, R. (2015). The reproducibility crisis in science: A statistical counterattack. *Significance* **12** 30–32.

PETERSEN, M. L., PORTER, K. E., GRUBER, S., WANG, Y. and VAN DER LAAN, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. *Stat. Methods Med. Res.* **21** 31–54. MR2867537 https://doi.org/10.1177/0962280210386207

PIRRACCHIO, R. and CARONE, M. (2018). 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.* **27** 2504–2518. MR3825922 https://doi.org/10.1177/0962280216682055

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974 https://doi.org/10.1093/biomet/70.1.41

ROTNITZKY, A., LI, L. and LI, X. (2010). A note on overadjustment in inverse probability weighted estimation. *Biometrika* **97** 997–1001. MR2746169 https://doi.org/10.1093/biomet/asq049

RUBIN, R. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Serv. Outcomes Res. Methodol.* **2** 169–88.

RUBIN, D. B. (2008). For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* **2** 808–804. MR2516795 https://doi.org/10.1214/08-AOAS187

SCHUEMIE, M. J., RYAN, P. B., HRIPCSAK, G., MADIGAN, D. and SUCHARD, M. A. (2018). Improving reproducibility by using high-throughput observational studies with empirical calibration. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **376** 1–17.

SHORTREED, S. M. and ERTEFAIE, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics* **73** 1111–1122. MR3744525 https://doi.org/10.1111/biom.12679

SMITH, G. C., SEAMAN, S. R., WOOD, A. M., ROYSTON, P. and WHITE, I. R. (2014). Correcting for optimistic prediction in small data sets. *Am. J. Epidemiol.* **180** 318–24. https://doi.org/10.1093/aje/kwu140

VELDKAMP, C. (2017). The human fallibility of scientists: Dealing with error and bias in academic research Ph.D. thesis.

WILLIAMS, R. J., TSE, T., HARLAN, W. R. and ZARIN, D. A. (2010). Registration of observational studies: Is it time? *CMAJ, Can. Med. Assoc. J.* **182** 1638–1642. https://doi.org/10.1503/cmaj.092252