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