Statistical Science
2020, Vol. 35, No. 3, 511–517
https://doi.org/10.1214/20-STS789
Main article: https://doi.org/10.1214/19-STS735

Rejoinder: A Nonparametric Superefficient Estimator of the Average Treatment Effect

David Benkeser, Weixin Cai and Mark J. van der Laan

1. INTRODUCTION

We thank each of the comment authors for their insights and perspectives on our work. The comments were wide-ranging in content and raised many interesting questions pertaining to our work and its place in the larger scope of research in the area. We address each commenter in turn.

2. LI

We thank Dr. Li for his interesting comment and novel proposal for stabilization in the context of estimating the average treatment effect. Li asks the question as to whether stabilization techniques that are common for inverse probability of treatment weighted (IPTW) estimators could stabilize doubly robust procedures in weakly identified settings. In essence, Li proposes to use a stabilized propensity score in combination with one-step estimation or TMLE. The stabilized propensity score is of the form $\bar{G}_0(w \mid h) = \bar{G}_0(w)/h(w)$, where $h : \mathcal{W} \to [0, 1]$ is some mapping that may depend on P_0. Several choices of h are discussed, such as

$$h(w) = \frac{\bar{G}_0(w)\{1 - \bar{G}_0(w)\}}{\int \bar{G}_0(w)\{1 - \bar{G}_0(w)\}\, dQ_{0,W}(w)}. \tag{1}$$

The author proposes a plug-in estimator h_n of h, based on an estimate of the propensity score, $\bar{G}_n$, and proceeds as usual with a one-step and TMLE procedure using the alternative propensity score estimator $\bar{G}_n(w \mid h_n) = \bar{G}_n(w)/h_n(w)$. The resultant estimators are found via simulation to have reasonable performance in the simulation settings considered in our paper.

Overall, Dr. Li's idea to bring in stabilization techniques from the IPTW literature to the doubly robust sphere is novel and interesting. However, we would like to highlight a potential difficulty when considering coupling this approach with machine learning or other nonparametric regression techniques. The potential problem is illustrated most directly by the analysis of Li's estimator in the case where $\bar{G}_0$ is known exactly, as in a stratified randomized trial. This setting is important, since it is one where asymptotically linear, doubly robust estimators can be generated under the weakest possible assumptions. We will argue that when the outcome regression is estimated nonparametrically Li's estimator may not achieve asymptotic linearity in even this "best-case" scenario.

Let $\psi^1_{n,*}$ be Li's TMLE of ψ^1_0, constructed based on the targeted outcome regression estimate $\bar{Q}^1_{n,*}$, the true stabilized propensity score $\bar{G}_0(\cdot \mid h)$ and the empirical distribution of W, $Q_{n,W}$. Below, we write Pf to denote $\int f(o)\, dP(o)$ for a given P-integrable function f and for each $P \in \mathcal{M}$. We also denote by P_n the empirical distribution function based on $O_1, \ldots, O_n$, so $P_n f = n^{-1} \sum_{i=1}^n f(O_i)$. We define $R_{0n} = P_0\{D^1(\cdot \mid \bar{Q}^1_{n,*}, Q_{n,W}, \bar{G}_0(\cdot \mid h)) - D^1(\cdot \mid \bar{Q}^1_{n,*}, Q_{n,W}, \bar{G}_0)\}$. A linearization of Ψ^1 along with straightforward algebra gives

$$\begin{aligned}
&\Psi^1(Q^1_{n,*}) - \Psi^1(Q^1_0) \\
&\quad = -P_0 D^1(\cdot \mid \bar{Q}^1_{n,*}, Q_{n,W}, \bar{G}_0) \\
&\quad = -P_0 D^1(\cdot \mid \bar{Q}^1_{n,*}, Q_{n,W}, \bar{G}_0(\cdot \mid h)) + R_{0n} \\
&\quad = (P_n - P_0) D^1(\cdot \mid \bar{Q}^1_{n,*}, Q_{n,W}, \bar{G}_0(\cdot \mid h)) + R_{0n},
\end{aligned} \tag{2}$$

where the third line follows since, by construction, the targeted estimate $\bar{Q}^1_{n,*}$ is such that $P_n D^1(\cdot \mid \bar{Q}^1_{n,*}, Q_{n,W}, \bar{G}_0(\cdot \mid h)) = 0$. The first term in the final equality is an empirical process and standard conditions can be assumed to control its behavior (Appendix B of the web supplement accompanying the original paper). However,

$$\begin{aligned}
R_{0n} &= \mathrm{E}_{P_0}\left(\{h(W) - 1\}\left[\frac{A}{G_0(W)}\{Y - \bar{Q}^1_n(W)\}\right]\right) \\
&= \mathrm{E}_{P_0}\left(\{h(W) - 1\}\right. \\
&\qquad \left. \times \left[\frac{A}{G_0(W)}\{\mathrm{E}_{P_0}(Y \mid A, W) - \bar{Q}^1_n(W)\}\right]\right) \\
&= \mathrm{E}_{P_0}\left(\{h(W) - 1\}\left[\frac{A}{G_0(W)}\{\bar{Q}^1_0(W) - \bar{Q}^1_n(W)\}\right]\right) \\
&= \mathrm{E}_{P_0}[\{h(W) - 1\}\{\bar{Q}^1_0(W) - \bar{Q}^1_n(W)\}].
\end{aligned}$$

In order for Li's estimator to be asymptotically linear with the claimed influence function, we would need to establish that $R_{0n} = o_{\mathrm{p}}(n^{-1/2})$. However, the form of R_{0n} is not second-order unless $h(w) = 1$ for all $w \in \mathcal{W}$ (in which

David Benkeser is an Assistant Professor of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia 30322, USA (e-mail: benkeser@emory.edu). Weixin Cai is a researcher at Citadel, Seattle, Washington USA. Mark J. van der Laan is Professor of Biostatistics and Statistics, University of California, Berkeley, California 94720, USA.

case Li's estimator reduces to a standard TMLE). Thus, in general we do not expect negligibility of this term.

What are the implications for inference? If $\bar{Q}_n$ is a maximum likelihood estimate based on a correctly specified parametric model, then $\psi^1_{n,*}$ will be asymptotically linear; *however*, there should be a first-order contribution to its influence function from R_{0n}. Thus, standard error estimates based on the variance of the efficient influence function alone may not be consistent for the true asymptotic variance of the estimator. Nevertheless, in this context, the nonparametric bootstrap should suffice to provide confidence intervals with valid coverage probability. On the other hand, if the outcome regression is estimated nonparametrically, we may expect nonstandard behavior of Li's estimator, since we cannot rule out the possibility that $n^{1/2}R_{0n}$ converges to $\pm\infty$. In particular, we expect that bias of $\psi^1_{n,*}$ may not converge faster than $n^{-1/2}$.

The situation is more difficult still when $\bar{G}_0$ is unknown since (2) is then replaced by $\Psi^1(Q^1_{n,*}) - \Psi^1(Q^1_0) = -P_0 D^1(\cdot \mid \bar{Q}^1_{n,*}, Q_{n,W}, \bar{G}_0) + R_{2,0n}$, where $R_{2,0n} = P_0[\{\bar{G}_n(\cdot \mid h_n) - \bar{G}_0\}/\bar{G}_n(\cdot \mid h_n)(\bar{Q}_n - \bar{Q}_0)]$. When the estimated propensity score targets the true propensity score, this term is second-order; otherwise, it will in general contribute to the first-order behavior of the estimator. The implications for inference in this setting are the same as above. For a more extensive analysis of $R_{2,0n}$ when the propensity estimator does not target the true propensity, see Benkeser et al. (2017).

2.1 Simulation Study

We provide a short simulation examining the phenomena described above. Our setting is intentionally simplistic. To simulate data, we drew W_1 from a Uniform(0, 1) distribution and independently drew W_2 from a Bernoulli(1/2). Given $W = w$, the treatment A was drawn from a Bernoulli distribution with $\bar{G}_0(w) = \text{logit}^{-1}(1 + 2w_1)$. Given $A = a, W = w$, the outcome was drawn from a Bernoulli distribution with $\bar{Q}^a_0(w) = \text{logit}^{-1}(-1 + w_1 - 2w_1w_2)$, for $a = 0, 1$. We studied three TMLE estimators of ψ^1_0: Li's proposed estimator based on the choice of h in (1), our CTMLE and a standard TMLE. We simulated 3000 datasets of size $n = 250, 500, 1000, 2000, 3000, 4000, 5000$ and compared estimators' Monte Carlo bias and their bias when scaled by $n^{1/2}$, a key property needed for asymptotic linearity. We plotted estimated sampling distributions of the centered estimators scaled by an oracle standard error (i.e., the Monte Carlo standard deviation over the 3000 data sets), as well as the estimated standard error based on the empirical standard deviation of the efficient influence function evaluated at each estimator's choice of propensity score estimator. Similarly, we studied coverage probability of nominal 95% confidence intervals based on oracle and estimated standard errors. We used the highly adaptive lasso (HAL, Benkeser and van der Laan, 2016, van der Laan, 2017) to estimate the outcome regression, propensity score and outcome-adaptive propensity score.

All estimators had small bias in large samples (Figure 1, left); however, the bias of Li's estimator is not unequivocally converging faster than $n^{-1/2}$ (Figure 1, right). On the other hand, the bias of CTMLE and TMLE appears to be converging at the proper rate. Evidence of this bias again appears in the sampling distribution of Li's estimator (Figure 2, top left), where we see that when scaled by an oracle standard error, the estimator has a small negative bias, even in large samples. However, the impact on confidence interval coverage is minimal (Figure 2, top right) In contrast, the most apparent feature of the sampling distribution of Li's estimator when scaled by an estimated standard error is that its variability is smaller than that of a standard Normal random variable, particularly in larger samples. This results in overcoverage of the confidence intervals based on the estimated standard error. In contrast, the CTMLE and TMLE have more standard asymptotic behavior with sampling distributions better approximating the asymptotic distribution in large samples and confidence intervals approach nominal coverage.

2.2 Concluding Thoughts

Li's idea of bringing stabilized propensity scores in doubly robust estimation is appealing, but there appear to be important theoretical considerations in cases of non

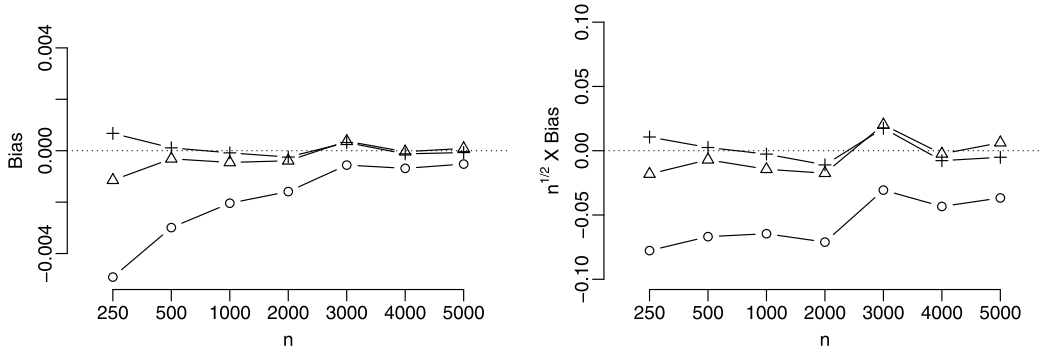

FIG. 1. *Bias (left) and $n^{1/2}$-scaled bias (right) of Li's estimator (circles), our proposed CTMLE (triangles), and a standard TMLE (+).*

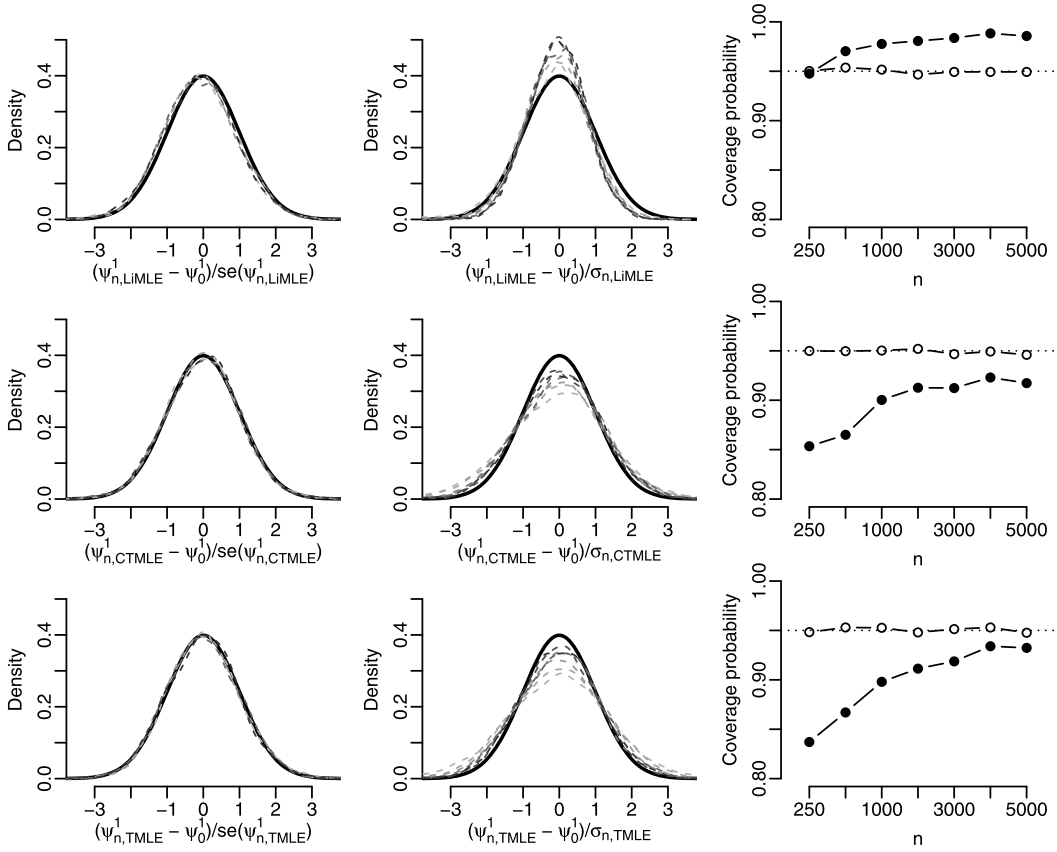

FIG. 2. *Sampling distribution of estimators centered and scaled by their true standard deviation* (*left*) *and their estimated standard error* (*middle*). *The estimators' distributions are shown in gray-scale, with darker color indicating larger sample sizes; a standard Normal distribution is shown in black. The right column shows coverage probability for a nominal* 95% *confidence interval based on estimated standard errors* (*black circles*) *or oracle standard errors* (*white circles*).

and semiparametric implementations. The analysis of Li's estimator highlights difficulties that can be expected in the context of doubly robust estimators when opting for nonstandard targets for the propensity score estimator. Our target propensity that conditions on the outcome regression was very carefully selected to avoid these difficulties. Nevertheless, it will be interesting in future research to better understand whether and how propensity score stabilization can be used in the context of flexible implementations of doubly robust estimators.

3. SCHNITZER

We offer our gratitude to Dr. Schnitzer for her comments on our work. Her perspectives on what is needed for wider adoption are quite welcome. In direct response to her call for more and better software, we have included an implementation of our proposed methodology in version 1.0.5 of the `drtmle` package for the `R` language (Benkeser, 2020). Beyond this, we add several comments.

Dr. Schnitzer's first point pertaining to model diagnostics points to an important gap in the literature. Beyond the lack of available software to facilitate diagnosis of positivity violations, there is still work to be done in understanding what diagnostics are most relevant for doubly robust estimators. While such diagnostics have been extensively studied for matching and IPTW estimators (e.g., various covariate balance metrics), there are fewer available studies identifying relevant diagnostics for doubly robust approaches. Moreover, there has not been a comprehensive evaluation of the impact that analytic decisions made based on these diagnostics may have on downstream inference. The open question then is to develop methodology that is able to appropriately diagnose when standard doubly robust estimators will struggle and adaptively shift estimation toward more stable versions of those estimators, and of course, as Dr. Schnitzer reminds us, to develop efficient and easy-to-use software that implements any such methods.

The second point discusses a need for more diverse software packages that can handle different study designs and estimation of different causal quantities. With this point, we could not agree more. Current trends in statistics and data science represent a sort of race, with the way in which data are collected leading the way. Methodology developments closely follow, while software development lags considerably behind. The reasons for this lag are likely multifaceted, including but not lim-

ited to: insufficient training in software development in doctoral programs, challenges of securing funding specifically for software development, and lack of emphasis on software development in traditional academic promotion processes. While there has recently been some improvement in these regards, we still have a ways to go.

Dr. Schnitzer's third point recommends improving computational run times associated with machine learning-based methodology. Certainly, this is a barrier to adoption of such methodology amongst analysts who are used to having analyses run in a matter of seconds. To a certain extent this increase in run time is unavoidable as we move toward more flexible analytical approaches. As the field moves toward more flexible approaches, it will become crucial to emphasize reproducible coding and unit tests so that the additional analytical time is primarily computer wall time rather than human time. The relative simplicity of many classical statistical analyses can allow bad coding practices to develop. After all, there is little issue with a program that crashes due to syntax error after running for half a second on a desktop. However (speaking from personal experience), a program crashing due to a syntax error after three days of run time on an expensive cloud computing unit is devastating. To help with the transition to more sophisticated analyses on larger data sets, we can emphasize best coding practices for reproducibility in our training of the next generation of statisticians and data scientists.

4. SHORTREED AND MOODIE

Drs. Shortreed and Moodie provide a timely and wide-ranging reflection on possible pitfalls of automation in data analysis. We thank them for their thoughtful contribution. In our response, we first clarify several points about our work before turning to the broader question of automation.

4.1 Points of Clarification

Quoting from Petersen et al. (2012), Drs. Shortreed and Moodie imply that our proposed procedure is "[in principle, settling] for a better estimate of a less interesting parameter" (Section 3). We do not believe this to be the case. The cited literature surrounding this statement discusses trimming propensity scores to estimate parameters that are easier to identify than the average treatment effect. However, we wish to emphasize that this is not the approach adopted in our work; we directly estimate average treatment effect. It is possible that Drs. Shortreed and Moodie are highlighting that the target propensity score of our procedure is less interesting than the true propensity score and we agree that this is probably true in many contexts. However, the goal of our work is not propensity score estimation as an end in itself, but rather as a means to the end of drawing stable inference on the average treatment effect.

The next two points of clarification pertain to the use of cross-validation-based ensemble learning, or super learning (van der Laan, Polley and Hubbard, 2007). Since this methodology was not discussed in the original paper, for the sake of completeness and to contextualize or responses below, we provide a brief description here. Super learning is a generalization of regression stacking (Wolpert, 1992, Breiman, 1996) and entails prespecifying a number of *candidate* estimators, each aimed at estimating the same target quantity. Cross-validation is used to determine an ensemble (e.g., convex combination) of the estimators that minimizes a cross-validated estimate of a user-specified risk criteria (e.g., mean squared-error). Oracle inequalities demonstrate that the resultant ensemble estimator has essentially the same or better asymptotic risk when compared to the best single estimator among all of the candidates. While super learning can be used outside of the context of TMLE (and vice versa), the two methods have often been combined in our past work.

Drs. Shortreed and Moodie state that, in the context of TMLE, super learning for estimation of a propensity score is "potentially harmful," while the same is not true for outcome regression estimation, which "[underscores] that the goals of prediction and causal inference can differ" (Section 2). We would like to expand this discussion. The cited works of Alam, Moodie and Stephens (2019) and Pirracchio and Carone (2018) deal specifically with propensity score matching, adjustment or weighting, and not doubly robust approaches. Indeed, a central assumption to the asymptotics of doubly robust approaches is $L^2(P_0)$-convergence of both the outcome regression and propensity score to their true respective counterparts. Thus, for doubly robust estimators the goals of prediction (e.g., having low mean squared-error) and causal inference seem to align quite well. On the other hand, in the context of the aforementioned "singly robust" approaches, standard implementations of treatment effect estimators are not expected to have standard asymptotic behavior when coupled with a standard implementation of super learning. For example, van der Laan (2014) provided an analysis of a super learner-based IPTW estimator and highlighted that additional effort is required to attain asymptotic linearity of such estimators. An alternative approach to attaining asymptotic linearity in this case is to utilize an undersmoothed minimum loss estimator (e.g., van der Laan, Benkeser and Cai, 2019). In either case, a careful application of super learning (or other machine learning approaches) is required to satisfy theoretical requirements in large samples. In small samples, various modifications have been shown to stabilize behavior in several contexts (e.g., propensity score truncation as in Bembom and van der Laan, 2008,

Xiao, Moodie and Abrahamowicz, 2013). It would be interesting to evaluate these modifications in the context of the extensive simulations performed by Alam, Moodie and Stephens (2019).

Overall, Alam, Moodie and Stephens (2019) rightfully point out that singly robust approaches are common in practice and provide an illustration of where naïve applications of super learning may not perform appreciably better than simple propensity score estimation approaches, such as main terms logistic regression. However, we do not believe that these results should dissuade practitioners from ever using more flexible propensity score estimators. Rather, the arguments of Alam, Moodie and Stephens (2019) point to a clear need for better communication of how such flexible methodologies can be appropriately employed in practice.

As a segue into our broader discussion of automation, we also wish to respond to Drs. Shortreed and Moodie on the point that "[when combined with super learning] TMLE is less automated than we might think." Based on our reading, we understand Drs. Shortreed and Moodie use *automated* to refer to methodology that is *mostly devoid of human input*. If this is indeed their intended definition, then we agree that super learning is not automated, nor should it be. In fact, rather than removing human input from modeling, the super learner framework should invite human collaboration. For example, in recent collaborations involving observational studies of influenza vaccine effectiveness, we required models for predicting influenza infection in a health care setting. With limited background knowledge of the biology of influenza, it would be extremely difficult to pose realistic parametric models, and thus we may be tempted to instead rely exclusively on black-box machine learning approaches, which require little subject matter expertise. Thankfully, our collaborators have been developing influenza prediction models for years and could anticipate where we should expect to see interactions between variables, which variables may have nonlinear relationships with the outcome, and other idiosyncrasies unique to influenza data. But naturally our collaborators had uncertainty about their models (e.g., should participant age be included as a linear term? a quadratic? in categories?). The super learner framework allows the past experiences of collaborators in influenza modeling to be incorporated in the analysis by including several different versions of the proposed regression models in the super learner, while providing an objective, prespecified means of making difficult modeling decisions. The appeal of the super learner framework is this ability to facilitate prespecified collaboration in the face of estimator uncertainty, rather than as a means of moving toward statistical methodology that requires no human input.

4.2 On Automation

We turn now to the broader question at the core of the comment by Drs. Shortreed and Moodie: should automation and data-driven analyses be preferred when inferential, rather than predictive, analyses are undertaken? We first focus on inferential questions that are confirmatory, as opposed to exploratory, in nature. By this we mean that there is an a-priori hypothesis that, for example, variable A has an effect on outcome B and we wish to use data to quantify the magnitude of this effect or test for the presence of such an effect. We consider exploratory analyses in the sequel.

We build our argument for where and when automation may be applied in the scientific process on a formal roadmap for inference (Petersen and van der Laan, 2014). The roadmap outlines the interplay of science and statistics en route to drawing conclusions from data. With this entire process in front of us, we can scrutinize areas where automation is useful and where it may lead us astray.

1. Specify knowledge about the system under study.
   - Using a structural causal model or related graphical technique, codify existing knowledge about how variables in the system under study do/do not causally relate to one another.
2. Specify observed data and link to causal model.
   - Determine the implications of the causal model on the observed data distribution. How does the sampling procedure relate to the causal model?
3. Specify target causal quantity.
   - Decide which variables in your system on which we would intervene in an "ideal experiment" and how we would intervene on those variables.
4. Assess identifiability of the causal parameter.
   - Given the assumptions made by the structural causal model and the sampling design, can we estimate the target causal quantity using observed data?
5. Commit to statistical model and estimand.
   - Select an estimand that is as close as possible to the target causal quantity given the potential limitations of the data.
6. Estimate the chosen statistical estimand.
   - Develop a prespecified analysis plan, possibly informed by a *blinded* analysis of the data. Execute the analysis plan on the real data.
7. Interpret the results.
   - Determine whether assumptions are sufficient to interpret results causally or as mere associations. Make explicit the assumptions under which the interpretation holds.

To begin, all statisticians would likely agree that automation should be desired in Step 6. Human intervention could and should occur in the development of an analysis plan (e.g., as in the influenza example above) and the development of robust code needed to execute that analysis. However, analytic decisions should ideally be made prior to unblinding of the data, while the actual data analysis should occur in a nearly fully automated way. Such automation prevents human bias from infecting the analysis and preserves interpretability of confidence intervals and hypothesis tests.

The next place where automation could play a limited role is in the interplay of Steps 3 and 4. Algorithms for identification of counterfactual distributions are available in the literature (among others, Tian and Pearl, 2002, Shpitser and Pearl, 2006) and can help enumerate which causal quantities are identifiable. Similarly, diagnostic procedures could be used to assess positivity assumptions that could lead to a lack of identifiability of some estimands. In both situations, it is still important to have a human in the loop to identify estimands that most appropriately answer the scientific question of interest and determine the most appropriate way forward. Or, if no suitable estimands can be identified, human input is required to identify what data should be collected in the future that would enable us to answer the question of interest.

It is interesting to note that the roadmap also highlights exactly where automation failed in the examples provided by Drs. Shortreed and Moodie. In the example of Google Flu, postmortem examinations revealed that, among other issues, the prediction algorithm failed to adapt to changes in the underlying Google search algorithms (failure of automation in Step 1) (Lazer et al., 2014). The system under study was changing over time and knowledge of this system was not incorporated into the architecture of predictions made by Google Flu. Conversely, for the example of the moon landing, we should all be grateful that Armstrong and Aldrin did indeed have a strong understanding of the system they were operating, such that human intervention was possible. The examples provided in predictive policing come down to inappropriate automation of Steps 2 and/or 3. Certainly, if bias is present in current policing data sets, we can expect that bias to propagate through to analyses derived from those data (failure of automation in Step 2). Moreover, recent research in the area has lead to the idea that "default" criteria used to train prediction algorithms (e.g., based purely on predictive accuracy) are not necessarily appropriate for the stated end (failure of automation in Step 3) (Corbett-Davies et al., 2017, Kusner et al., 2017). Finally, the study of health plan disenrollment and suicide risk is another example of the failure of automation in Step 2, where the observed data were inappropriately linked to the causal model.

In summary, in the setting of confirmatory analysis, we believe that opportunities for full automation are, at least for the time being, rather limited. However, automation is absolutely essential in the estimation stage in order to deliver accurate, unbiased and reproducible inferences. Looking to the future, further automation of the scientific process will likely require new artificial intelligence technologies that are capable of more appropriately adjudicating cause and effect (i.e., better automation of Steps 1 and 2, Hartnett, 2018).

We turn now to exploratory data analysis, where it is more natural to "let the data ask the questions." In some settings, data are quite rich, but we lack a basic understanding of what the interesting questions might be. For example, in the intensive care unit, myriad measurements are made on patients in real time: blood pressure monitoring, heart rate monitoring, oxygen saturation, level of intravenous drug administration, etc. Some health outcomes, such as sepsis, are so poorly understood that collaborators often ask open-ended questions such as, "Can we use the data to determine what variables are important in this setting?" In these instances, science has not progressed far enough to provide testable hypotheses; nevertheless, we should like to learn *something* from such a rich source of data. At the very least, we can hope to generate hypotheses that inform the next generation of confirmatory studies. This has motivated our developments pertaining to *data-adaptive target parameters*, which provide a formal framework for using data to learn what questions may be interesting to ask, while simultaneously providing an answer to that question in the form of an estimated association or effect (van der Laan and Luedtke, 2015, Hubbard, Kherad-Pajouh and van der Laan, 2016, Hubbard, Kennedy and van der Laan, 2018). However, even this process requires a human in the loop to provide the class of questions from which the "most interesting" questions are chosen based on the data.

To conclude, in the modern Big Data era and at a time when artificial intelligence/machine learning are at a crest of enthusiasm, Drs. Shortreed and Moodie pose important questions as to how much of the scientific process can reliably be turned over to automation. In our opinion, for the time being, the answer is relatively little. Nevertheless, automation at the point of execution of a prespecified analysis plan is fundamental to the very notion of frequentist statistics and must be emphasized to draw robust and reproducible scientific conclusions from data.

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

BEMBOM, O. and VAN DER LAAN, M. J. (2008). Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators. U.C. Berkeley Division of Biostatistics Working Paper Series, Working Paper 230.

BENKESER, D. (2020). drtmle: Doubly-robust nonparametric estimation and inference. R package version 1.0.5. https://doi.org/10.5281/zenodo.844836

BENKESER, D. and VAN DER LAAN, M. J. (2016). The highly adaptive lasso estimator. In *Proceedings of the International Conference on Data Science and Advanced Analytics* 2016 689–696. https://doi.org/10.1109/DSAA.2016.93

BENKESER, D., CARONE, M., VAN DER LAAN, M. J. and GILBERT, P. B. (2017). Doubly robust nonparametric inference on the average treatment effect. *Biometrika* **104** 863–880. MR3737309 https://doi.org/10.1093/biomet/asx053

BREIMAN, L. (1996). Stacked regressions. *Mach. Learn.* **24** 49–64. https://doi.org/10.1007/bf00117832

CORBETT-DAVIES, S., PIERSON, E., FELLER, A., GOEL, S. and HUQ, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the* 23*rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *KDD* '17 797–806. Association for Computing Machinery, New York. https://doi.org/10.1145/3097983.3098095

HARTNETT, K. (2018). To build truly intelligent machines, teach them cause and effect [online; accessed 13-January-2020].

HUBBARD, A. E., KENNEDY, C. J. and VAN DER LAAN, M. J. (2018). Data-adaptive target parameters. In *Targeted Learning in Data Science*. *Springer Ser. Statist*. 125–142. Springer, Cham. MR3820724

HUBBARD, A. E., KHERAD-PAJOUH, S. and VAN DER LAAN, M. J. (2016). Statistical inference for data adaptive target parameters. *Int. J. Biostat.* **12** 3–19. MR3505683 https://doi.org/10.1515/ijb-2015-0013

KUSNER, M. J., LOFTUS, J., RUSSELL, C. and SILVA, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems* 30 (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds.) 4066–4076. Curran Associates, Red Hook, NY.

LAZER, D., KENNEDY, R., KING, G. and VESPIGNANI, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science* **343** 1203–1205.

PETERSEN, M. L. and VAN DER LAAN, M. J. (2014). Causal models and learning from data: Integrating causal modeling and statistical estimation. *Epidemiology* **25** 418–426. https://doi.org/10.1097/EDE.0000000000000078

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

SHPITSER, I. and PEARL, J. (2006). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence* **2** 1219. AAAI Press, Menlo Park, CA; MIT Press, Cambridge, MA.

TIAN, J. and PEARL, J. (2002). A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence* 567–573.

VAN DER LAAN, M. J., BENKESER, D. and CAI, W. (2019). Causal inference based on undersmoothing the highly adaptive lasso. In *Proceedings of the* 2019 *AAAI Spring Symposium*. Association for the Advancement of Artificial Intelligence, Menlo Park, CA.

VAN DER LAAN, M. J. and LUEDTKE, A. R. (2015). Targeted learning of the mean outcome under an optimal dynamic treatment rule. *J. Causal Inference* **3** 61–95.

VAN DER LAAN, M. J. (2014). Targeted estimation of nuisance parameters to obtain valid statistical inference. *Int. J. Biostat.* **10** 29–57. MR3208072 https://doi.org/10.1515/ijb-2012-0038

VAN DER LAAN, M. (2017). A generally efficient targeted minimum loss based estimator based on the highly adaptive Lasso. *Int. J. Biostat.* **13** Art. ID 20150097. MR3724476 https://doi.org/10.1515/ijb-2015-0097

VAN DER LAAN, M. J., POLLEY, E. C. and HUBBARD, A. E. (2007). Super learner. *Stat. Appl. Genet. Mol. Biol.* **6** Art. ID 25. MR2349918 https://doi.org/10.2202/1544-6115.1309

WOLPERT, D. H. (1992). Stacked generalization. *Neural Netw.* **5** 241–259. https://doi.org/10.1016/s0893-6080(05)80023-1

XIAO, Y., MOODIE, E. E. and ABRAHAMOWICZ, M. (2013). Comparison of approaches to weight truncation for marginal structural Cox models. *Epidemiol. Methods* **2** 1–20.