Statistical Science
2020, Vol. 35, No. 3, 472–475
https://doi.org/10.1214/20-STS771
Main article: https://doi.org/10.1214/19-STS728

Comment: Outcome-Wide Individualized Treatment Strategies

Ashkan Ertefaie and Brent A. Johnson

1. INTRODUCTION

We congratulate VanderWeele, Mathur and Chen (VMC) for their timely and interesting contribution to the emerging field of longitudinal designs for observational studies. Their paper provides intriguing guidelines on evaluating robustness and sensitivity to potential unmeasured confounding in outcome-wide studies. We expand on their discussion and point out another important application of the outcome-wide studies in developing individualized treatment strategies.

2. OVERT AND HIDDEN BIASES

The primary concern in causal inference is bias that does not diminish as the sample size increases. In general, there are two types of biases: overt and hidden. An overt bias is one that can be seen in the data at hand, for example, the imbalance of a measured pre-treatment covariate across the treatment groups. A hidden bias is similar to an overt bias in the sense that both are caused by the imbalance across treatment groups but the former cannot be seen in the available data because the required information was not observed or recorded (Rosenbaum, 2002). One of the core assumptions in causal inference is the no unmeasured confounder assumption which rules out the presence of hidden biases. But even under the no unmeasured confounder assumption, there is no guarantee that the causal inference methods can produce unbiased estimates, in general. Bias can still manifest itself through certain model misspecification.

Assuming that there is no hidden bias, data adaptive techniques can provide powerful tools to reduce the chance of model misspecification thereby reducing the bias in the treatment effect estimation. However, these methods may lead to an estimator with an unknown asymptotic behavior because of slower rate of convergence than root-n rate. For example, an inverse probability weighting estimator is no longer asymptotically linear if the propensity scores are estimated using a data adaptive technique (e.g., random forest (Liaw, Wiener et al., 2002)). This is because the convergence rate of the inverse probability weighting estimator entirely depends on the rate of convergence of the postulated model for the propensity score. Double robust estimators are alternatives that can overcome this shortcoming. Double robust estimators are based on modeling both the propensity score and the outcome processes and are consistent for the target parameter of interest when any one of two models is consistently estimated. The asymptotic linearity of the double robust estimators is guaranteed when both nuisance parameters are consistently estimated with convergence rate faster than $n^{1/4}$ (Van der Laan and Robins, 2003). Although double robust models facilitate the use of data-adaptive techniques for modeling the nuisance parameters, the resulting estimator can be irregular with large bias and slow rate of convergence when one of the nuisance parameters is inconsistently estimated. Undersmoothing and targeting techniques have been proposed to mitigate this issue (van der Laan, 2014, Benkeser et al., 2017, van der Laan, Benkeser and Cai, 2019).

As pointed out by VMC, the analysis results can be considerably biased by investigator choice after looking at the data. This is even more of a concern when data adaptive techniques are used for two reasons. First, these methods often involve multiple tuning parameters that have to be specified using the data. Often, investigators decide to set some of the tuning parameters to the default (i.e., prespecified) values and choose the others using cross-validation. Second, there are many data adaptive techniques that could be used and the concluding results may depend on the method used. These can be a source of bias if the decision is made after seeing the results. Outcome-wide studies can mitigate this problem if the investigator uses the same modeling and tuning approaches for all the outcomes included in the analyses.

Ensemble learning methods (e.g., super learner) seems to be particularly helpful in reducing the chance of bias caused by model misspecification and researcher bias. Ensemble learning methods combine different user specified data-adaptive techniques (e.g., random forest, generalized additive models, support vector regression) in an optimal way to produce a predictive model which is superior to each individual algorithm included in the ensemble learning model (van der Laan, Polley and Hubbard, 2007).

Ashkan Ertefaie is an Assistant Professor, Department of Biostatistics and Computational Biology, University of Rochester, Rochester, New York 14642, USA (e-mail: ashkan_ertefaie@urmc.rochester.edu). Brent A. Johnson is a Professor, Department of Biostatistics and Computational Biology, University of Rochester, Rochester, New York 14642, USA (e-mail: brent_johnson@urmc.rochester.edu).

Specifically, they first build a predictive model for each algorithm and then uses cross-validation to find the optimal weighted combination of the predicted values as a final output. The library of the methods included in the ensemble learning must be specified prior to seeing the results. The size of the library of methods can grow as a polynomial rate of the sample size and can include different tuning parameter specifications of a certain method, thereby resolving both concerns raised above (Van Der Laan and Dudoit, 2003).

3. SENSITIVITY ANALYSIS FOR HIDDEN BIASES

Drawing causal inference from observational studies based on regression or propensity score analyses rely on the untestable assumption that there are no unmeasured confounders. In general, because it is impossible to be certain about this assumption, the causal interpretation of the estimated treatment effect is questionable. There are alternative methods that can be used to obtain unbiased treatment effect estimates even in the presence of unmeasured confounders at the expense of relying on some other untestable assumptions.

A sensitivity analysis is a specific statement about the magnitude of hidden bias that would need to be present to explain the associations actually observed in a particular study. Weak associations in small (or large) studies can be explained away by very small biases, but only a very large bias can explain a strong association in a large study. Let D be binary treatment variable such that $D \in \{0, 1\}$ and $\boldsymbol{X}$ be a vector of pretreatment variables. Also, let $\pi(\boldsymbol{x}) = Pr(D = 1 \mid \boldsymbol{X} = \boldsymbol{x})$ denote the propensity score. There is hidden bias if two units with the same observed covariates $\boldsymbol{x}$ have different chances of assignment to treatment. Let $\pi_j/(1-\pi_j)$ and $\pi_k/(1-\pi_k)$ be the odds that units j and k receive the treatment. Suppose we knew that this odds ratio for units with the same $\boldsymbol{x}$ was at most some number $\Gamma \geq 1$,

$$\frac{1}{\Gamma} \leq \frac{\pi_j(1-\pi_k)}{\pi_k(1-\pi_j)} \leq \Gamma \quad \text{for all } j \text{ and } k \text{ with } \boldsymbol{x}_j = \boldsymbol{x}_k. \tag{1}$$

In this formulation, $\Gamma = 1$ denotes a study free of hidden bias and $\Gamma > 1$ denotes a study with certain magnitude of hidden bias. Thus, a possibly complex unmeasured confounding pattern can be summarized using a scalar value Γ. Rosenbaum proposed series of sensitivity analysis tools based on the odds ratio in (1) that can be used in variety of matching techniques (Rosenbaum, 1987, 1988, 1989, 1991, 2002). To perform the sensitivity analyses, one starts with $\Gamma = 1$ and incrementally increases Γ and calculates the corresponding testing p-values for each Γ until the type-I error rate is reached. Then that particular Γ value represents the level of sensitivity of the results to unmeasured confounding. Confidence intervals corresponding to each Γ value can also be constructed. A $(1-\alpha)$ sensitivity interval for a causal parameter with sensitivity parameter is a random interval that in at least $(1-\alpha)$ of studies will contain the true parameter value assuming that the true sensitivity parameter Γ_0 satisfies $\Gamma_0 < \Gamma$. VMC proposed an alternative measure of sensitivity to hidden bias called the E-value. For a binary outcome Y,

$$\text{E-value} = \text{RR}_{\text{obs}} + \sqrt{\text{RR}_{\text{obs}}(\text{RR}_{\text{obs}} - 1)},$$

where $\text{RR}_{\text{obs}} = \Pr(Y = 1 \mid D = 1, \boldsymbol{x})/\Pr(Y = 1 \mid D = 0, \boldsymbol{x})$. The E-value can be interpreted as "the minimum strength of association on the risk ratio scale that an unmeasured confounder would need to have with both the treatment and the outcome to fully explain away a specific treatment-outcome association, conditional on the measured covariates" (VanderWeele and Ding, 2017). One advantage of the sensitivity analyses proposed by Rosenbaum is that it can be done completely nonparametrically regardless of the dimension of the measured confounders (i.e., $\boldsymbol{X}$). For example, we can perform matching using rank based Mahalanobis distance and then construct a confidence interval by inverting a rank based permutation test (Rosenbaum, 1989, 2002). The E-value results, however, may vary as a function of the postulated conditional outcome model given the treatment and measured confounders. This can complicate the interpretation of the E-value particularly when the dimension of $\boldsymbol{X}$ is moderate to large. However, unlike Rosenbaum's approach where the sensitivity parameter Γ only reflects the effect of an unmeasured confounder on the propensity score, the E-value considers the effect an unmeasured confounder on both the outcome and the treatment models. This suggests that the E-value may carry more information about the actual strength of an unmeasured confounder on the estimated effect than Rosenbaum's approach.

One important limitation of the sensitivity analysis is that there is no guideline on what values of the sensitivity parameter are deemed small indicating that the residual confounding is a serious threat. In a study that researchers are confident that they have included most of important confounders in the analysis, a relatively small sensitivity parameter value may be considered implausible while a study that does not include several potentially important confounder may be considered sensitive to hidden bias even if the sensitivity parameter value is larger than the former study. This highlights the importance of carefully-designed observational studies. Indeed, sensitivity analysis is useful for quantifying how much bias would be needed to change the conclusions of the study, but it can never prove that there is a treatment effect because we do not know for certain how much hidden bias there could be

and all effect estimates are sensitive when the the bias is sufficiently large.

Outcome-wide designs can provide valuable insight on implausible sensitivity values for a given study. Suppose there is an outcome known to be unaffected by the treatment denoted as $Y^\dagger$. Then we first perform the sensitivity analysis using the primary outcome Y and suppose we obtain that sensitivity parameter value of Γ^* is sufficient to explain away the detected causal effect. Second, we test whether the null hypothesis of no effect on $Y^\dagger$ holds true under the hidden bias with magnitude of Γ^*. If $\Gamma = \Gamma^*$ is rejected, the point of greatest sensitivity to bias has been rejected and we can conclude that Γ^* is an implausible value. If $\Gamma = \Gamma^*$ is not rejected, then the outcome $Y^\dagger$ does not help in reducing our concern about the hidden bias of Γ^* (see Section 6.2 of Rosenbaum, 2002).

4. OUTCOME-WIDE DESIGN IN INDIVIDUALIZED TREATMENT STRATEGIES

Some treatments do not affect all the subjects the same way. The main goal of individualized treatment strategies (ITS) is to use individual patient characteristics to inform a personalized treatment plan that leads to the best healthcare possible for each patient (Zhang et al., 2012, Zhao, Small and Ertefaie, 2017, Ertefaie and Strawderman, 2019). An optimal ITS is the one that optimizes an outcome of interest. The current literature in ITS is mostly focused on developing methods to estimate an optimal ITS based on a single outcome, with some attention given to the case of multiple outcomes (Laber, Lizotte and Ferguson, 2014, Fard and Pineau, 2009, Lizotte and Laber, 2016). In real life, clinical decision-making aims to balance several potentially competing outcomes (e.g., symptom relief, clinician's qualitative assessment, side effects, costs, patient's preference).

Outcome-wide design provides a unique opportunity to consider multiple outcomes in developing optimal decision rules. Specifically, one can construct and report a set of best treatment options with respect to each outcome for a given patient using the multiple comparison with the best method (Hsu, 1981, 1984). The set is constructed such that it contains the true best treatment option with certain probability (e.g., 0.95) (Ertefaie et al., 2016). Let $\mathcal{K}_j(\boldsymbol{x})$ correspond to the set of best treatments with respect to outcome Y_j, $j = 1, 2, \ldots, K$, for a patient with baseline covariates $\boldsymbol{x}$. For each patient, clinicians can then aggregate the constructed sets of best treatments over multiple outcomes to from an outcome-wide set of best treatments denoted as $\mathcal{B}(\boldsymbol{x})$. If the sets overlap across all the outcomes, then $\mathcal{B}(\boldsymbol{x})$ will include the intersection of those sets. If the sets do not overlap, then we will report the union of $\mathcal{K}_j(\boldsymbol{x})$ for $j = 1, 2, \ldots, K$ as the set of best $\mathcal{B}(\boldsymbol{x})$ and leave the tie breaking to the physician and patient preference.

The quality of the estimated optimal ITS can be seriously affected in the presence of overt biases caused by model misspecification and hidden biases. The former can be mitigated by using double robust estimators that leverage data adaptive techniques to estimate the required nuisance parameters (Zhang et al., 2012). Sensitivity analysis approaches could potentially be used to assess the sensitivity of the suggested optimal treatment for a given patient to different magnitude of hidden biases. Under treatment effect heterogeneity, subjects may benefit from different treatment options depending on their baseline characteristics. Also, among those who benefit from a certain treatment, the magnitude of the treatment effect may vary. As a result, the sensitivity level of the optimal treatment choice may vary substantially from patient to patient. This can potentially complicate the interpretation of the sensitivity analysis results. In settings with a binary treatment option and a single outcome, there are subjects who benefit significantly from either of the treatments and there are subjects whose treatment effect is not significantly different across the two treatment arms. As the sensitivity parameter Γ increases, the number of patients in the indifference zone will increase. Thus, for example, one metric to report could be the proportion of patients who significantly benefit from a single treatment as a function of Γ. However, the sensitivity analysis can get complicated in the outcome-wide design settings and merits further research.

5. LITTLE VERSUS BIG "M" IN MULTIPLE OUTCOMES

First, given where the statistics field is in the analysis of massive data sets, it seems that a central question to the analysis of multiple outcomes is how large is large in the word "multiple"? In VMC, the authors define the vector of multiple outcomes as $(Y_1, \ldots, Y_K)$, and so our question rephrased in the author's notation is how large is K? Methods that might be ideal for the analysis of 10 outcomes may be unsuitable for the analysis of 10 million outcomes. As VMC suggest in Section 8.1, the very name ''outcome wide" analysis is intended to draw connections to genome wide association studies (GWAS) where investigators consider millions of genetic variants. However, in the motivating data example, VMC analyze $K = 24$ outcomes which might suggest the scale of K does not align with GWAS studies and that other methods might be better suited to analyze multiple outcome in the proposed setting when K is moderate.

Second, while the analysis of multivariate Gaussian outcomes serves as a principal cornerstone in the analysis of multiple outcomes, there are now well-established statistical methods for modeling jointly multiple outcomes of mixed types (e.g., copulas) and these other approaches

could provide additional insight into the data that the current outcome-wide analysis framework proposed by VMC does not address. In addition to potential efficiency gains that VMC state in Section 2.7, statisticians model jointly multiple outcomes to estimate and draw inference on the associations among those outcomes. In the motivating example, there may be a scientific interest in the association between measures of emotional or psychological well-being and adverse health behaviors and possibly whether this association is mediated by other factors. Developing such a method may or may not be a simple combination of existing tools but would connect the proposed framework with an important aspect of the analysis of multiple outcomes research.

REFERENCES

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

ERTEFAIE, A. and STRAWDERMAN, R. L. (2019). Robust Q-learning Technical report. Available at arXiv:2003.12427.

ERTEFAIE, A., WU, T., LYNCH, K. G. and NAHUM-SHANI, I. (2016). Identifying a set that contains the best dynamic treatment regimes. *Biostatistics* **17** 135–148. MR3449856 https://doi.org/10.1093/biostatistics/kxv025

FARD, M. M. and PINEAU, J. (2009). MDPs with non-deterministic policies. In *Advances in Neural Information Processing Systems* 1065–1072.

HSU, J. C. (1981). Simultaneous confidence intervals for all distances from the "best". *Ann. Statist.* **9** 1026–1034. MR0628758

HSU, J. C. (1984). Constrained simultaneous confidence intervals for multiple comparisons with the best. *Ann. Statist.* **12** 1136–1144. MR0751303 https://doi.org/10.1214/aos/1176346732

LABER, E. B., LIZOTTE, D. J. and FERGUSON, B. (2014). Set-valued dynamic treatment regimes for competing outcomes. *Biometrics* **70** 53–61. MR3251666 https://doi.org/10.1111/biom.12132

LIAW, A., WIENER, M. et al. (2002). Classification and regression by randomForest. *R News* **2** 18–22.

LIZOTTE, D. J. and LABER, E. B. (2016). Multi-objective Markov decision processes for data-driven decision support. *J. Mach. Learn. Res.* **17** Art. ID 211. MR3595145

ROSENBAUM, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* **74** 13–26. MR0885915 https://doi.org/10.1093/biomet/74.1.13

ROSENBAUM, P. R. (1988). Sensitivity analysis for matching with multiple controls. *Biometrika* **75** 577–581. MR0967598 https://doi.org/10.1093/biomet/75.3.577

ROSENBAUM, P. R. (1989). On permutation tests for hidden biases in observational studies: An application of Holley's inequality to the Savage lattice. *Ann. Statist.* **17** 643–653. MR0994256 https://doi.org/10.1214/aos/1176347131

ROSENBAUM, P. R. (1991). Sensitivity analysis for matched case-control studies. *Biometrics* **47** 87–100. MR1108691 https://doi.org/10.2307/2532498

ROSENBAUM, P. R. (2002). Overt bias in observational studies. In *Observational Studies* 71–104. Springer, Berlin.

VAN DER LAAN, M. J. and DUDOIT, S. (2003). Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. U.C. Berkeley Division of Biostatistics Working Paper Series 130.

VANDERWEELE, T. J. and DING, P. (2017). Sensitivity analysis in observational research: Introducing the E-value. *Ann. Intern. Med.* **167** 268–274. https://doi.org/10.7326/M16-2607

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. J., BENKESER, D. and CAI, W. (2019). Efficient estimation of pathwise differentiable target parameters with the undersmoothed highly adaptive lasso. Preprint. Available at arXiv:1908.05607.

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

VAN DER LAAN, M. J. and ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer, Berlin.

ZHANG, B., TSIATIS, A. A., LABER, E. B. and DAVIDIAN, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics* **68** 1010–1018. MR3040007 https://doi.org/10.1111/j.1541-0420.2012.01763.x

ZHAO, Q., SMALL, D. S. and ERTEFAIE, A. (2017). Selective inference for effect modification via the lasso. Preprint. Available at arXiv:1705.08020.