Statistical Science
2020, Vol. 35, No. 3, 479–483
https://doi.org/10.1214/20-STS791
Main article: https://doi.org/10.1214/19-STS728
© Institute of Mathematical Statistics, 2020

Rejoinder: The Future of Outcome-Wide Studies

Tyler J. VanderWeele, Maya B. Mathur and Ying Chen

We thank Daniel (2020), Vansteelandt and Dukes (2020) and Ertefaie and Johnson (2020) for their thoughtful, insightful and enlightening commentaries on our paper (VanderWeele, Mathur and Chen, 2020). We have learned a great deal from their comments, discussion and proposals. The outcome-wide approach is still in its infancy and, as pointed out by the commentators, there are certainly numerous ways to refine and extend what we had proposed as a basic template. The analytic approaches even to estimate causal effects of a single time-fixed exposure on a single subsequent outcome have increased dramatically over the past decades. The range of considerations and decisions that arise when considering multiple outcomes are yet more vast, and thus, over time, there may likewise be an array of principled analytic options for outcome-wide studies, too. We suspect that many of the seeds for that potentially vast array of options are likely to be found in the commentaries of Daniel (2020), Vansteelandt and Dukes (2020) and Ertefaie and Johnson (2020). We will respond to their various remarks by considering how they give rise to important cautions, important extensions and important alternatives to the practice of outcome-wide studies.

IMPORTANT CAUTIONS

The commentators raise a range of important points and caveats to the implementation and interpretation of the outcome-wide analytic approach that we proposed, some of which indeed may not have received due attention in our paper. Daniel (2020) rightly points out that when, with covariate data that is contemporaneous with the exposure, it is unclear whether a particular covariate is a confounder or a mediator, the approach of considering analyses both with and without the covariate will not necessarily bound the causal effect. The two analyses can be biased in the same direction when the covariate's effect on the exposure is of the opposite direction of the exposure's effect on the covariate. While we do still think considering both analyses with and without the covariate is valuable, we certainly acknowledge that concordance of these two analyses does not necessarily indicate a clear conclusion. Discordance should raise cause for concern; but even with concordance one should be cautious in interpretation. When the temporal structure of the data is such that covariate levels prior to the exposure cannot be adjusted for, and it is thus unclear whether a covariate is a confounder or a mediator, we think it will often be difficult to draw causal conclusions (VanderWeele, 2015).

Daniel (2020) also rightly points out the potential importance, if one is using multiple imputation to handle missing data, of including all outcomes simultaneously in imputation models. In our early work implementing the outcome-wide approach, we had indeed neglected this (Chen and VanderWeele, 2018; Chen et al., 2019a, 2019b), but in all of our more recent empirical outcome-wide analyses (Chen et al., 2019c, Chen, Kubzansky and VanderWeele, 2019, Long et al., 2020, Kim et al., 2020), this is indeed how we have proceeded. The analytic approach to these outcome-wide studies is certainly still evolving, and there will almost certainly be other refinements to it, like this one, a point to which we will also return below.

Daniel (2020) raises further important concerns about potential positivity violations. The analysis, regardless of whether using propensity scores, or regression models, or doubly-robust methods, requires that the groups with and without exposure have overlap in the covariates. If certain exposure or treatment decisions are made deterministically then this can be violated. While this is indeed a well-known fact within causal inference, there was arguably not sufficient emphasis of it in our paper, and as Daniel (2020) rightly notes, the problem may be compounded by the large number of covariates for which adjustment might be made in an outcome-wide approach. Checking adequate covariate overlap can often be facilitated by estimating propensity scores, an alternative analytic approach for outcome-wide studies discussed in our paper, and also advocated for in the outcome-wide context by Vansteelandt and Dukes (2020), and which we will also consider further below.

Tyler J. VanderWeele is John L. Loeb and Frances Lehman Loeb Professor of Epidemiology, Harvard T.H. Chan School of Public Health, and Institute for Quantitative Social Science, Harvard University, Boston, Massachusetts 02115, United States (e-mail: tvanderw@hsph.harvard.edu). Maya B. Mathur is Assistant Professor, Quantitative Sciences Unit, Stanford University, Palo Alto, California 94304, United States. Ying Chen is Research Associate, Institute for Quantitative Social Science, Harvard University, Cambridge, Massachusetts 02138, United States.

Vansteelandt and Dukes (2020) also rightly express a concern with outcome-wide analyses that investigators, when examining many outcomes, may be less careful about attempting to measure all of the predictors of each and every outcome, which may be problematic when it comes to confounding control. While they say this point arises from their discussion of their own proposed three-step approach, it is certainly a concern with outcome-wide studies more generally, regardless of the analytic approach taken, as indeed indicated in our paper. Regardless of the analytic approach chosen, investigators should try to collect data on all pre-exposure predictors of any of the outcomes under study. This will of course be more challenging with many outcomes, and there can be temptation to neglect this undertaking. However, if the importance of this issue is kept in mind, investigators can work toward addressing it. The outcome-wide approach inevitably sacrifices some depth for breadth, but efforts can be made to minimize the trade-off and sacrifice. We would never want to minimize the importance of the depth allowed by specific, more narrowly focused, studies examining a single exposure and a single outcome. There will always be room for such work, as indeed suggested by Daniel (2020) and noted in our paper. However, when the outcome-wide approach is used, it will be best to do whatever is possible to minimize the potential loss of rigor and detail that is inevitably, to at least a certain extent, entailed.

IMPORTANT EXTENSIONS

In our paper, we considered a number of possible extensions of the outcome-wide approach including to interaction outcome-wide studies, outcome-wide studies for time-varying exposures, quasi-experimental outcome-wide studies, lagged exposure-wide studies, and mediator- and moderator-wide studies. One intriguing extension that we did not consider was raised by Ertefaie and Johnson (2020) concerning the use of outcome-wide studies in individualized treatment strategies. We think this is an incredibly important and fruitful, yet likely difficult, area for future methodological development, and one that extends well beyond any of the considerations in our paper. There has been recent important work on statistical approaches to optimal individualized treatment selection for a single outcome (e.g., Luedtke and van der Laan, 2015, 2016; VanderWeele et al., 2019). This itself is a challenging problem, risks of overfitting are substantial, and recent work suggests that principled approaches require a relatively large sample size before their asymptotic optimality properties set in (Luedtke, Sadikova and Kessler, 2019). These challenges are likely to be compounded when considering multiple outcomes. And yet, as Ertefaie and Johnson (2020), rightly note, this is in fact the actual decision-making framework that is needed in many patient contexts. Ertefaie and Johnson (2020) explicitly mention considerations of symptom relief, clinicians' qualitative assessment, side effects, costs and patients' preferences. However, in major treatment decisions that potentially involve life-long side effects, the outcomes of interest may involve trade-offs between physical health and life expectancy on the one hand, and happiness, purpose and sense of mastery through the capacity to work, quality of social relationships and one's personal financial situation on the other (VanderWeele, McNeely and Koh, 2019, VanderWeele, 2017). Current medical practice tends to prioritize and optimize life-expectancy and years of disease-free survival but in treatment decisions concerning tongue cancer or bladder cancer, other aspects of flourishing may be as important, or more important, to patients (VanderWeele et al., 2019). Methodology to incorporate multiple outcomes into optimal individualized treatment decision-making would be of tremendous value in this regard, though, as noted by Ertefaie and Johnson (2020), will certainly be challenging statistically.

Ertefaie and Johnson (2020) also raise yet another important opportunity for the extension of, or perhaps better said, expansion of, outcome-wide studies and that concerns settings with a very large number of outcomes. The example given in our paper had 24 outcomes, and indeed all of our extant empirical outcome-wide papers have fewer than 100 outcomes. We fully admit that our paper on outcome-wide methodology, and the practical advice given therein, was shaped entirely around a setting in which there were numerous outcomes, but not in the thousands, and certainly not in the millions! We always had in view settings in which the number of subjects vastly exceed the number of outcomes. In settings in which this is not so, we believe that the analytic approaches would likely have to be altered considerably. Our reporting guidelines would certainly fall apart. Moreover, the very matter of interpretation of results of outcome-wide studies with a very large number of outcomes would become extremely challenging. A patient might reasonably weigh the effects of ten, or perhaps even fifty, relevant outcomes, but it would be difficult even to convey, and much more so to reason about, thousands or millions of outcomes. In cases of extreme sparsity in which the exposure is thought to affect only a few of very many outcomes, interpretation might again simplify. One might imagine examining the effects of a slight change in a water supply system on a vast range of plant and animal life, expecting that only very few, if any, would be substantially affected. However, outside of such settings of sparsity, these issues of interpretation, in addition to analysis, will raise difficulties and considerations that again extend well beyond our paper. These possible extensions would, however, be interesting, and perhaps important, to pursue in future work.

IMPORTANT ALTERNATIVES

Our paper laid out a number of the conceptual and design considerations for outcome-wide longitudinal studies. It also suggested, and illustrated, a particular statistical approach to the analysis for such outcome-wide studies and briefly discussed other possible analytic strategies or variations. As noted above, the analytic approach that we are using in our own empirical work on outcome-wide studies has evolved, and will likely continue to evolve. We were therefore intrigued by a number of the proposals put forward by Vansteelandt and Dukes (2020).

In light of the false-positive considerations that can arise with multiple testing under model misspecification, Vansteelandt and Dukes (2020) recommend that priority be given in outcome-wide studies to propensity score approaches. The concerns about positivity raised by Daniel (2020) lend further weight to this as an analytic option. We are not opposed to the use of propensity scores in outcome-wide studies, though our paper did mention a few caveats. There is the further concern, noted by Vansteelandt and Dukes (2020) that, with regard to potential model misspecification in outcome wide studies, this may have the problem of "all-your-eggs-in-one-basket" wherein one is "betting on one propensity score model being correct," rather than "spreading the risk of misspecification over different postulated outcome models."

Vansteelandt and Dukes (2020) address this concern by proposing a three-step approach wherein a model is also specified for each outcome and for the exposure, both for the purposes of covariate selection and then later to obtain effect estimates. The approach of Vansteelandt and Dukes (2020) allows for different covariates for each outcome, which, they rightly argue, has some advantages. Vansteelandt and Dukes (2020) claim that their proposed procedure also addresses concerns about "an investigator fitting a series of regressions and choosing one to their liking." While an honest application of their proposed three-step approach is indeed possible, there are dangers. In allowing different covariates in each outcome model, even if this is done by an automated procedure, the additional "investigator degree-of-freedom" (Simmons, Nelson and Simonsohn, 2011, Gelman and Loken, 2014) of choosing the tuning parameter may be introduced, as noted also by Ertefaie and Johnson (2020). With each different outcome, it may indeed seem reasonable that each would have a different tuning parameter, but this again gives the investigator more opportunities to pick and choose among the analyses. We do not have especially strong views on this matter. It is never possible to completely eliminate investigator degrees of freedom and evaluation of evidence will always depend in part on investigator integrity. We suspect that which approach is best in practice will require further consideration and refinement over time, both as to how problematic differing tuning parameter choices across outcomes are in practice, and with regard to the sample size requirements needed for the data-adaptive selection procedures to have good finite-sample properties. We would not be surprised if there are indeed a class of settings in which Vansteelandt and Dukes' three-step approach is advantageous in practice. That space still needs to be clearly delineated, but again, as noted above, we believe that the analytic approach to outcome-wide studies will indeed continue to evolve over time.

Concerns over multiple testing are of course another important issue in outcome-wide studies and we devoted an entire section of our paper to it. It is on this analytic aspect of outcome-wide studies that we suspect our views and practices will likely change the most as time passes. The range of contexts in which our new multiple testing metrics (Mathur and VanderWeele, 2018) provide additional insight is unclear, and unfortunately we have not yet been able to extend that approach to binary and count outcomes. Vansteelandt and Dukes (2020) introduce another alternative approach to both global and individual outcome testing. While the theoretical properties of their proposed approach are indeed intriguing, we are concerned that it relies too much on a hypothesis testing framework that dichotomizes evidence. As we hope we made clear in our commentary, and as per other recent discussion (Greenland, 2017, Amrhein, Greenland and McShane, 2019), we believe that the dichotomization of evidence within science (reject the null versus not) has been highly problematic.

Evidence should be viewed more on a continuum. We believe that in the interpretation of that evidence in outcome-wide studies there need to be clear indications as to how an investigator or reader may be additionally misled by examining evidence for multiple relationships. However, even the Bonferroni adjusted threshold we certainly do not view as a definitive means to draw a dichotomized conclusion, but rather as one way, amongst many, to scale the evidence to help see the impact that examining multiple relationships simultaneously may have had. Vansteelandt and Dukes (2020) point out that a drawback of their approach is that it may imply a large number of false positives whenever the first test falsely rejects. We think this is a very important drawback that also illustrates the problematic nature of a dichotomized hypothesis testing approach. We prefer approaches to multiple testing (i.e., to the examining of multiple relationships at once) that more easily allow investigators to see the continuous nature of the evidence under consideration. We do not have definitive answers here as to how best to do this, and believe that this is an important, and still very much open, area of inquiry as to how to best carry this out, both within a paper, and also across papers for a given topic. As noted in our paper, we believe evidence accumulates across studies, and that often conclusions can only be more definitively drawn by combining evidence across studies either by meta-analysis, or careful description of the evidence, or other approaches. It is often problematic to try to draw definitive conclusions from a single study.

As noted above, and as made clear yet further by the critiques and proposals of Vansteelandt and Dukes (2020), there is likely considerable room for alternative approaches and improvements of outcome-wide studies both with respect to statistical modeling and with respect to issues of multiple testing.

Ertefaie and Johnson (2020) also raise alternative approaches to sensitivity analysis to unmeasured confounding in outcome-wide studies. We had proposed reporting E-values (VanderWeele and Ding, 2017) for each outcome examined in order to assess the minimum strength of association on the risk ratio scale that an unmeasured confounder would have to have with both the exposure and the outcome, conditional on the measured covariates, to completely explain away the observed exposure-outcome relationship. Ertefaie and Johnson (2020) discuss the advantages and disadvantages of using the E-value as compared with a sensitivity analysis parameterization of Rosenbaum (2002) that focuses exclusively on the treatment assignment selection. We have likewise discussed similarities and differences of the E-value as contrasted with other sensitivity analysis parameterization elsewhere (VanderWeele, Ding and Mathur, 2019). While we are deeply committed to sensitivity analysis in observational causal inference, and thus in outcome-wide studies, we are in no way committed to the specific use of the E-value for carrying out such sensitivity analysis. There are many very good sensitivity analysis techniques available that help assess the sensitivity or robustness of effect estimates to potential unmeasured confounding (Rosenbaum, 2002, Rothman, Greenland and Lash, 2008, Lash, Fox and Fink, 2009). The E-value was introduced as a particularly simple way to assess, and report, sensitivity to unmeasured confounding for those who may not be willing to use more involved, or more thorough, approaches (VanderWeele and Ding, 2017, VanderWeele, Mathur and Ding, 2019). In general, however, a more extensive sensitivity analysis will be more informative, and the E-value is not a fully adequate substitute. However, in the context of outcome-wide studies, an advantage of the E-value is that, because it is effectively a single value, it is simple to report across a wide-range of outcomes. When thus employed, however, one again has a sacrifice in outcome-wide studies of some depth for greater breadth.

Ertefaie and Johnson (2020) note that an important limitation of sensitivity analysis is that there is often no guideline on what values of the sensitivity analysis parameters are deemed small or for definitively determining when residual confounding is a serious threat. This is indeed so. In other work (VanderWeele and Mathur, 2020), we have suggested a move towards better practices in reporting E-values by reporting associations of all measured covariates with the outcome in a multivariate regression model. This begins to help assess, for a specific outcome, what magnitudes of association might be considered "large." In the context of outcome-wide analyses with many outcomes, such reporting for each outcome would require a large amount of space. This might still be possible in a very lengthy online supplement. However, an alternative might be a single table wherein, for each outcome, the three largest associations with the outcome (after inversion for protective associations) across covariates are reported. If the study included a very large number of covariates, then in addition to the three largest associations it might also be of interest to report, say, the top 1st, 5th and 10th quantiles of association magnitude (again, after inversion for protective associations). Such information could again be helpful in assessing what magnitude of associations with each outcome might be considered "large," and thus the extent to which observed exposure-outcome associations might be plausibly sensitive or robust to potential unmeasured confounding.

THE FUTURE OF OUTCOME-WIDE STUDIES

As noted in the Introduction, the outcome-wide approach is still in its infancy. As made clear in the three commentaries (Daniel, 2020, Vansteelandt and Dukes, 2020, Ertefaie and Johnson, 2020) and the discussion here, there are numerous opportunities for refinement and improvement, and for extension to new settings. We described a basic analytic template for such studies that we have been using in our own empirical work. As evidence relevant to shaping best practices for outcome-wide studies emerges, our recommended analytic template for these studies will likely shift over time. Many of the considerations of Daniel (2020), Vansteelandt and Dukes (2020) and Ertefaie and Johnson (2020) already point to possible refinements that might be developed further. Nevertheless, we believe that the value and importance of the outcome-wide approach, viewed as a design, rather than as an analytic template, will persist. Regardless of how best practices for the analysis of outcome-wide studies evolve, its value and advantages in allowing the publication of null results, minimizing investigator degrees of freedom, facilitating comparison of effect sizes and allowing for a more efficient and rapid expansion of knowledge will still be present, almost irrespective of how analytic approaches to such studies may change. We look forward to future developments and investigations. As in our paper, we believe that the use of such outcome-wide designs will contribute to an accelerated and more accurate advancement of knowledge and, if a broad range of outcomes is examined, to the promotion of a fuller human flourishing.

REFERENCES

Amrhein, V., Greenland, S. and McShane, B. (2019). Scientists rise up against statistical significance. *Nature* **567** 305–307. https://doi.org/10.1038/d41586-019-00857-9

Chen, Y., Kubzansky, L. D. and VanderWeele, T. J. (2019). Parental warmth and flourishing in mid-life. *Soc. Sci. Med.* **220** 65–72.

Chen, Y. and VanderWeele, T. J. (2018). Associations of religious upbringing with subsequent health and well-being from adolescence to young adulthood: An outcome-wide analysis. *Am. J. Epidemiol.* **187** 2355–2364. https://doi.org/10.1093/aje/kwy142

Chen, Y., Kim, E. S., Koh, H. K., Frazier, A. L. and VanderWeele, T. J. (2019a). Sense of mission and subsequent health and well-being among young adults: An outcome-wide analysis. *Am. J. Epidemiol.* **188** 664–673.

Chen, Y., Harris, S. K., Worthington, E. L. Jr. and VanderWeele, T. J. (2019b). Religiously or spiritually-motivated forgiveness and subsequent health and well-being among young adults: An outcome-wide analysis. *J. Posit. Psychol.* **14** 649–658. https://doi.org/10.1080/17439760.2018.1519591

Chen, Y., Haines, J., Charlton, B. M. and VanderWeele, T. J. (2019c). Positive parenting improves multiple aspects of health and well-being in young adulthood. *Nat. Hum. Behav.* **3** 684–691.

Daniel, R. (2020). A new template for empirical studies: From positivity to Positivity. *Statist. Sci.* **35** 476–478.

Ertefaie, A. and Johnson, B. A. (2020). Comment: Outcome-wide individualized treatment strategies. *Statist. Sci.* **35** 472–475.

Gelman, A. and Loken, E. (2014). The statistical crisis in science. *Am. Sci.* **102** 460–465.

Greenland, S. (2017). Invited commentary: The need for cognitive science in methodology. *Am. J. Epidemiol.* **186** 639–645. https://doi.org/10.1093/aje/kwx259

Kim, E. S., Whillans, A. V., Lee, M. T., Chen, Y. and VanderWeele, T. J. (2020). Volunteering and subsequent health and well-being in older adults: An outcome-wide longitudinal approach. *Am. J. Prev. Med.* **59** 176–186.

Lash, T. L., Fox, M. P. and Fink, A. K. (2009). *Applying Quantitative Bias Analysis to Epidemiologic Data*. Springer, New York.

Long, K. N., Kim, E. S., Chen, Y., Wilson, M. F., Worthington, E. L. Jr and VanderWeele, T. J. (2020). The role of hope in subsequent health and well-being for older adults: An outcome-wide longitudinal approach. *Global Epidemiol.* **2** 100018.

Luedtke, A., Sadikova, E. and Kessler, R. C. (2019). Sample size requirements for multivariate models to predict between-patient differences in best treatments of major depressive disorder. *Clin. Psychol. Sci.* **7** 445–461.

Luedtke, A. R. and van der Laan, M. J. (2015). Targeted learning of the mean outcome under an optimal dynamic treatment rule. *J. Causal Inference* **3** 61–95.

Luedtke, A. R. and van der Laan, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Statist.* **44** 713–742. MR3476615 https://doi.org/10.1214/15-AOS1384

Mathur, M. B. and VanderWeele, T. J. (2018). New metrics for multiple testing with correlated outcomes. Preprint. https://doi.org/10.31219/osf.io/k9g3b

Rosenbaum, P. R. (2002). *Observational Studies*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR1899138 https://doi.org/10.1007/978-1-4757-3692-2

Rothman, K. J., Greenland, S. and Lash, T. L. (2008). *Modern Epidemiology*. Lippincott Williams & Wilkins, Philadelphia, PA. 345–380.

Simmons, J. P., Nelson, L. D. and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22** 1359–1366.

VanderWeele, T. J. (2015). *Explanation in Causal Inference*: *Methods for Mediation and Interaction*. Oxford Univ. Press, New York.

VanderWeele, T. J. (2017). On the promotion of human flourishing. *Proc. Natl. Acad. Sci. USA* **31** 8148–8156.

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

VanderWeele, T. J., Ding, P. and Mathur, M. B. (2019). Technical considerations in the use of the E-value. *J. Causal Inference* **7** 1–11. https://doi.org/doi.org/10.1515/jci-2018-0007

VanderWeele, T. J. and Mathur, M. B. (2020). Commentary: Developing best practice guidelines for the reporting of E-values. *Int. J. Epidemiol.* https://doi.org/10.1093/ije/dyaa094

VanderWeele, T. J., Mathur, M. B. and Chen, Y. (2020). Outcome-wide longitudinal designs for causal inference: A new template for empirical studies. *Statist. Sci.* **35** 437–466.

VanderWeele, T. J., Mathur, M. B. and Ding, P. (2019). Correcting misinterpretations of the E-value. *Ann. Intern. Med.* **170** 131–132.

VanderWeele, T. J., McNeely, E. and Koh, H. K. (2019). Reimagining health: Flourishing. *J. Am. Med. Assoc.* **321** 1667–1668.

VanderWeele, T. J., Luedtke, A. R., van der Laan, M. J. and Kessler, R. C. (2019). Selecting optimal subgroups for treatment using many covariates. *Epidemiology* **30** 334–341. https://doi.org/10.1097/EDE.0000000000000991

Vansteelandt, S. and Dukes, O. (2020). On the potential for misuse of outcome-wide study designs, and ways to prevent it. *Statist. Sci.* To appear.