Submitted to Statistical Science

On the potential for misuse of outcome-wide study designs, and ways to prevent it

Stijn Vansteelandt and Oliver Dukes

Ghent University and the London School of Hygiene and Tropical Medicine

We congratulate the authors, VanderWeele, T.J., Mathur, M.B. and Chen, Y. (2020) (hereafter referred to as VMC), for making an interesting and important proposal, and thank the Editor for the opportunity to comment on it. We agree with VMC that outcome-wide epidemiology has the potential to overcome many of the weaknesses of the traditional epidemiological approach. Scientific reports that express the effects of exposure on a variety of different outcomes provide a more complete view on the exposure impact, while lessening the risk of selective analysis and reporting. We see much value in it, though caution is warranted. In this commentary, we highlight a number of key limitations, which will in turn suggest preferred analysis strategies that we find important to consider in addition to (or instead of) those described by VMC.

1. BIAS INFLATION

With the analysis of multiple outcomes comes a growing of risk of bias in the effect of the exposure on (at least one of) those outcomes. Such inflated risk of bias may be the result of the more elaborate need for modelling (e.g. modelling each outcome separately) and the ensuing risk of model misspecification, the increased risk of (informative) missing data in those outcomes, a potentially reduced lack of care in collecting data on risk factors for all these outcomes (see Section 3) or in modelling the outcomes' dependence on measured risk factors, ... This expresses itself in particular into an inflated risk of Type I errors. Such inflation is not acknowledged by multiplicity adjustments such as the Bonferroni correction, which assume the absence of bias.

Figure 1 about here.

To appreciate this, let $\hat{\theta}_j$ express the estimated effect of exposure on the jth outcome ($j = 1, ..., k$). Suppose that $\hat{\theta}_j$ is normally distributed around θ_j with standard deviation $\sigma/\sqrt{n}$, where n is the sample size. Suppose further that the exposure has no effect on any of the outcomes, but that θ_j is nonetheless normally distributed with mean θ and standard deviation τ, which may both differ from zero as a result of bias. Under the above settings, the probability to find the exposure being associated with at least one of k mutually independent outcomes

Department of Applied Mathematics, Computer Sciences and Statistics, Ghent University Krijgslaan 281, S9, 9000 Gent, Belgium(e-mail: stijn.vansteelandt@ugent.be; oliver.dukes@ugent.be)

at the $\alpha 100\%$ significance level, when Bonferroni correction is used, equals

$$1 - \left[\Phi\left\{ \frac{-\Phi^{-1}(\alpha/2k) - \theta\sqrt{n}/\sigma}{\sqrt{1+\tau^2 n/\sigma^2}} \right\} - \Phi\left\{ \frac{\Phi^{-1}(\alpha/2k) - \theta\sqrt{n}/\sigma}{\sqrt{1+\tau^2 n/\sigma^2}} \right\} \right]^k.$$

Figure 1 displays this for $n = 100, \sigma = 1, \alpha = 0.05$ and $\theta = 0, \tau = 0.1$ (left), amounting to bias up to 2 standard errors away from zero for most outcomes, $\theta = 0.1, \tau = 0$ (middle), amounting to bias of 1 standard error for all outcomes, and $\theta = 0.1, \tau = 0.1$ (right), amounting to bias between -1 and 3 standard errors away from zero for most outcomes. These figures visualise the growing risk of false detections that may result from an accumulated risk of bias across all outcomes.

In view of these concerns, it is essential in our opinion that outcome-wide epidemiologic analyses be based on propensity scores. Since the same propensity score model can be used across all analyses, analyses that solely rely on correct specification of a propensity score model (see Sections 2 and 3 for specific proposals) do not suffer an increasing risk of model misspecification bias as more outcomes are being considered. In particular, their risk of bias due to model misspecification is the same as in the traditional epidemiologic design, in which one primary outcome is carefully studied. Further support for a propensity score analysis comes when drawing a parallel with outcome-wide randomised experiments; here, the propensity score is known by design, rendering an analysis that solely relies on correct specification of a propensity score (model) arguably the method of choice. For similar reasons that confounding bias in outcome-wide epidemiologic designs is - in our opinion - best addressed using propensity scores, outcome missingness due to dropout (in which case all outcomes are missing) is best addressed using analyses that solely rely on correct specification of a dropout model. This prevents an increasing risk of bias as more outcomes are being considered, as the same model can then apply to all outcomes.

Betting on one propensity score model being correct may also pose an increased risk (when that model is misspecified) as opposed to spreading the risk of misspecification over different postulated outcome models. In our opinion this need not be the case, however, as the need for more modelling may also imply a reduced care in building these models. In spite of this, in Section 3 we will propose a strategy which inherits the above mentioned advantages of a propensity score analysis, while not betting entirely on correct specification of a propensity score model.

2. ANALYSIS OF OUTCOME-WIDE VERSUS TRADITIONAL DESIGNS

It is instructive to contrast the outcome-wide epidemiologic design with the traditional epidemiologic design in which one primary outcome is carefully studied. This shows that the consideration of multiple outcomes, in the way proposed by VMC, may imply dilution of evidence when some of those outcomes are not or only indirectly affected by the exposure (e.g. via its effect on previously considered outcomes). Figure 2 (bottom, right) illustrates this for a setting where 1 (upper, left), 2 (upper, right), 3 (bottom, left) and 10 (bottom, right) of the considered outcomes are affected by the exposure. In particular, we consider mutually independent and normally distributed effect estimates $\hat{\theta}_j, j = 1, ..., k$ with standard deviation $\sigma/\sqrt{n}$, of which $l < k$ have mean $\theta \neq 0$ and the others have

mean zero. Figure 2 displays the probability to find the exposure being associated with at least one of the k outcomes at the $\alpha 100\%$ significance level, when separate tests with Bonferroni correction are used (solid, black). The solid black lines in Figure 2 were calculated as

$$1 - \left[\Phi\left\{-\Phi^{-1}(\alpha/2k) - \theta\sqrt{n}/\sigma\right\} - \Phi\left\{\Phi^{-1}(\alpha/2k) - \theta\sqrt{n}/\sigma\right\}\right]^l (1-\alpha/k)^{k-l},$$

for $\theta = 0.5, \sigma = 1, n = 10$ and $\alpha = 0.05$. Note the effect of dilution in all 4 panels as additional outcomes are considered that are not affected by the exposure. We view this as undesirable as it implies a potential loss of power, relative to the traditional epidemiologic design.

It seems tempting to prevent such power loss by first conducting a global test whether any of the outcomes is impacted by the exposure. Such global test may for instance be based on propensity scores via the likelihood ratio test whether $\theta_1 = ... = \theta_k = 0$ in model

$$\text{logit}P(A = 1|Z, Y_1, ..., Y_k) = \beta' Z + \sum_{j=1}^{k} \theta_j Y_j,$$

where A is the dichotomous exposure of interest and Z is a set of variables that is sufficient to adjust for confounding of the exposure effect on all outcomes $Y_j, j = 1, ..., k$. The red dashed lines in Figure 2 show that also the global test dilutes evidence as redundant outcomes are being added. Nonetheless, major power gains can sometimes be achieved relative to the strategy proposed by VMC, and contrary to what is somewhat suggested by VMC, who expect the increase in efficiency via global inference to be "modest". The red lines were calculated as the probability that a non-central chi-square distributed random variable with k degrees of freedom and non-centrality parameter $l\theta^2 n/\sigma^2$ exceeds the $(1-\alpha)100\%$ percentile of the central chi-square distribution with k degrees of freedom.

Figure 2 about here.

It follows from the above discussion that a global test can help prevent some of the power loss, which the proposal by VMC may suffer relative to a traditional epidemiologic design. It may even imply an increasing potential to detect outcomes that are impacted by the exposure. For example, several individual tests on the threshold of statistical significance could result in a (highly) significant global p-value, giving researchers the incentive to increase the number of outcomes in the analysis. We also view this as undesirable, as we believe the purpose of the outcome-wide epidemiologic design should not be to artificially increase power via the consideration of multiple outcomes.

We therefore recommend that the analysis of the outcome-wide epidemiologic design proceeds as follows. First, we test whether at least one of the outcomes is impacted by the exposure. For this, we perform a global test (e.g., the above suggested test), which returns a p-value p_{global}. We moreover perform each of the k unadjusted individual tests obtained by testing whether $\theta_j = 0$ in model

$$\text{logit}P(A = 1|Z, Y_j) = \beta' Z + \theta_j Y_j,$$

which returns a p-value p_j. We then propose to reject the null hypothesis that none of the outcomes is impacted by the exposure at the $\alpha 100\%$ significance level

only when

$$\max\left(p_{\text{global}}, \min_j p_j\right) < \alpha,$$

for the following reasons. By using p_{global}, we ensure protection of the family-wise error rate (at level α) of the overall procedure. By using the maximum of p_{global} and $\min_j p_j$, we moreover prevent an artificial increase of power via the consideration of multiple outcomes, as this first test can only do as good as the best of the individual tests. In doing so, note that our use of unadjusted p-values p_j prevents a power loss relative to the traditional epidemiologic design (and is justified via the earlier reliance on p_{global}).

Next, when this first test is significant, one may proceed to evaluate the individual tests to assess precisely which outcomes are affected by the exposure. Having already protected the family-wise error rate, this in principle necessitates no further multiplicity adjustment (when, as noted by the associate editor, one can content oneself with weak control of the family-wise error rate). Use of the unadjusted p-values is also attractive as it delivers a subsequent procedure that is consistent with the first: by taking the maximum of p_{global} and $\min_j p_j$ we ensure that when it is smaller than α (so that the test rejects the global null hypothesis), then at least one of the individual tests will also reject the corresponding null hypothesis at the $\alpha 100\%$ significance level. This is in contrast to common post-hoc procedures. One drawback is that it may imply a large number of false positives whenever the first test falsely rejects. Whenever this is a concern, one may instead proceed by rejecting the null hypothesis corresponding to the individual test $j^* = \text{argmin}_j p_j$ that resulted in the smallest p-value, in order to be consistent with the global test, and then adjusting the cut-off with which the remaining p-values are contrasted. In the Appendix, we explain how this can be done with the aim of controlling the expected number of remaining hypotheses that are falsely rejected whenever the first test rejects and $\theta_1 = ... = \theta_k = 0$, or the probability that at least one of the remaining hypotheses is falsely rejected whenever the overall test rejects and $\theta_1 = ... = \theta_k = 0$. While we provide a decision procedure in the Appendix, it remains to be studied how to construct corresponding adjusted p-values and confidence intervals, and whether strong control of the family-wise error rate is achievable along similar lines.

3. CONFOUNDER SELECTION

VMC recognise that in most epidemiologic analyses there is uncertainty surrounding which variables should be adjusted for in order to control confounding. The additional challenge in outcome-wide analyses is that one must identify the confounders of every exposure-outcome relationship considered. VMC also recognise that selecting an adjustment set through fitting a series of models for each Y_j and choosing only those that are strongly correlated with the outcome is problematic. Leeb, H. and Pötscher, B. M. (2005) show that for a single exposure and outcome, such procedures can lead to biased exposure effect estimators with complex non-normal distributions. In light of the previous concerns about bias inflation with an increasing number of outcomes, we emphasise the need for a principled approach to inference after confounder selection in outcome-wide epidemiology.

The preference of the authors is to adjust for a common set of covariates across all analyses. One could separately assess which covariates are predictive of each of the outcomes (adjusting for the exposure) using stepwise strategies or penalisation methods such as the Lasso; the same could be done to check which covariates are predictive of the exposure. Only variables that are not associated with the exposure or any of the outcomes are then excluded. Such a proposal is closely related to "double selection," recently proposed in the economics literature (Belloni, A., Chernozhukov, V. and Hansen, C., 2014; Belloni, A., Chernozhukov, V. and Wei, Y., 2016). Remarkably, under certain assumptions double selection procedures deliver hypothesis tests and confidence intervals that are *uniformly valid*, essentially meaning that there is a minimal sample size at which they attain their nominal size/coverage (within certain error margins), no matter what the data-generating process is. This counters the commonly-held wisdom that post-selection inferences do not honestly reflect the uncertainty induced via the data-adaptive selection procedure. Nevertheless, VMC also recognise that such a procedure could lead to a very large adjustment set in outcome-wide analyses.

We therefore suggest the following procedure; first, postulate a regression model for the conditional mean of the exposure A:

$$E(A|Z) = g^{-1}(\tau + \delta' Z)$$

where $g(.)$ is a known link function. When A is binary, it is typical to choose $g(x) = \text{logit}(x)$, but our proposal generalises to other choices of $g(x)$. We will first select variables associated with the exposure, e.g. via fitting the above model with a Lasso penalty (or using stepwise variable selection); let $\hat{P}$ denote the estimates of $E(A|Z)$ from a model refitted using only the selected covariates. Secondly, for outcome Y_1, we consider the linear model

$$E(Y_1|A, Z) = \alpha_1 + \beta_1 A + \gamma_1' Z.$$

We will select variables associated with the outcome (conditional on the exposue) e.g. via fitting the above model with a Lasso penalty forcing the exposure into the model. Let $Z^{(1)}$ refer to the vector of selected variables in this step. In the third step we fit the linear model

$$E(Y_1|A, Z) = \alpha_1 + \beta_1 A + \beta_{P_1} \hat{P} + \gamma_1^{*'} Z^{(1)}$$

and test whether β_1 differs from 0. One can obtain a p-value via standard software, so long as a sandwich estimator of the standard error is used. Steps 2 and 3 can then be repeated for $(Y_2, ..., Y_k)$. This procedure can also be extended to test the global null hypothesis that $\beta_1 = \beta_2 = ... = \beta_k = 0$ by fitting the third step regressions together using software for multiple-equation estimation e.g. the packages for generalised method of moments estimation available in Stata and R (Chaussé, P., 2010).

Such a proposal has the advantage that it is expected to deliver uniformly valid p-values and confidence intervals, similar to double selection (see the justification in Farrell, M.H. (2015) and Dukes, O. and Vansteelandt, S. (2019)). Although the authors object to allowing the adjustment variables $Z^{(j)}$ to depend on the choice of outcome j, the above procedure delivers the same set of covariates in the exposure model used in each analyses. Their concerns about investigators

fitting a series of regressions and choosing one to their liking are also addressed. It is indeed tempting to drop covariates that are weakly associated with Y_j but strongly with A. Doing so may result in a more precise estimate of the exposure effect (and potentially in a lower p-value), but also one that is biased, given that such variables may be important confounders. However, in our proposal, such variables will be nevertheless flagged via adjustment for $\hat{P}$ in the third step. We have focused on linear models for Y_j here, but the procedure extends to count or binary outcomes (see Dukes, O. and Vansteelandt, S. (2019) for details). The above proposal can in principle be adapted to deliver valid results when the propensity score model is misspecified, provided that the outcome model is correct (see the discussion at the end of Section 1). However, modifications of the fitting strategy may then be required to retain uniform validity in the presence of variable selection (Dukes, O. and Vansteelandt, S., 2019; Dukes, O., Avagyan V. and Vansteelandt, S., 2019).

The discussion above also has implications for study design and data-collection. A concern about outcome-wide epidemiology is that with a single outcome, the investigators may make more effort in collecting data on predictors of the exposure and outcome of interest. With multiple outcomes, it is easier to be less careful about measuring the predictors of all outcomes, which reduces the potential for valid confounding adjustment in case one had not realised a certain outcome predictor to be also predictive of the exposure. Therefore, in collecting data, for each Y_j considered one should ideally try to collect as many variables that predict *either* the exposure *or* the outcome as possible, to improve the chances of measuring all confounders.

APPENDIX A: POST-HOC TESTING PROCEDURE

Building on Branson, Z. and Bind, M.-A. (2019), we recommend a randomisation procedure whereby first a propensity score model for $P(A = 1|Z)$ is fitted. This model is then used to randomly reassign treatment to all participants, thereby imposing the null hypothesis that $\theta_1 = ... = \theta_k = 0$. For the mth rerandomised data, $m = 1, ..., M$ with M e.g. 10000, we perform the global test, $p^{(m)}_{\text{global}}$, as well as each of the k unadjusted individual tests, $p^{(m)}_j$, and calculate $p^{(m)} = \max\left(p^{(m)}_{\text{global}}, \min_j p^{(m)}_j\right)$. Over all repetitions where $p^{(m)} < \alpha$ for some chosen significance level α, e.g. 0.05, we then exclude the most significant test $j^{*(m)} = \text{argmin}_j p^{(m)}_j$ and evaluate either $q_e(\alpha^*)$, the expected number of remaining tests with $p^{(m)}_j < \alpha^*$, or $q_p(\alpha^*)$, the chance that at least one of the remaining tests has $p^{(m)}_j < \alpha^*$. We then determine the value α^* for which either $q_e(\alpha^*)$ or $q_p(\alpha^*)$ takes a pre-specified level, e.g. 0.05. The choice between a procedure based on $q_e(\alpha^*)$ or $q_p(\alpha^*)$ should ideally be guided by context; the use of $q_p(\alpha^*)$ is expected to be more conservative. The uncertainty in the estimated propensity scores can be taken into account by drawing the propensity score coefficients from their sampling distribution in each of the repetitions $m = 1, ..., M$.

ACKNOWLEDGEMENT

The authors gratefully acknowledge support from FWO Grants G.016116.N and 1S05916N, and BOF Grant BOF.24Y.2017.0004.01 and BOF.01P08419.

REFERENCES

VanderWeele, T.J. and Mathur, M.B. and Chen, Y. (2020). Outcome-wide longitudinal designs for causal inference: a new template for empirical studies. *Statistical Science* **??**, ??–??.

Leeb, H. and Pötscher, B. M. (2005). Model Selection and Inference: Facts and Fiction. *Econometric Theory* **21**, 21–59.

Chaussé P.(2010). Computing Generalized Method of Moments and Generalized Empirical Likelihood with R. *Journal of Statistical Software* **34**.

Farrell, M.H (2015). Robust inference on average treatment effects with possibly more covariates than observations *Journal of Econometrics* **189**, 1–23.

Dukes, O. and Vansteelandt, S. (2019). How to obtain valid tests and confidence intervals after propensity score variable selection? *Statistical Methods in Medical Research* , doi: 10.1177/0962280219862005.

Dukes, O. and Avagyan, V. and Vansteelandt, S. (2019). High-dimensional doubly robust tests for regression parameters *ArXiv*, arXiv:1805.06714; forthcoming in Biometrics.

Branson, Z. and Bind, M.-A. (2019). Model Selection and Inference: Facts and Fiction. *Statistical Methods in Medical Research* **28**, 1378–1398.

Belloni, A. and Chernozhukov, V. and Hansen, C. (2014). Inference on Treatment Effects after Selection among High-Dimensional Controls. *The Review of Economic Studies* **81**, 608–650.

Belloni, A. and Chernozhukov, V. and Wei, Y. (2016). Post-Selection Inference for Generalized Linear Models With Many Controls. *Journal of Business & Economic Statistics* **34**, 606–619.

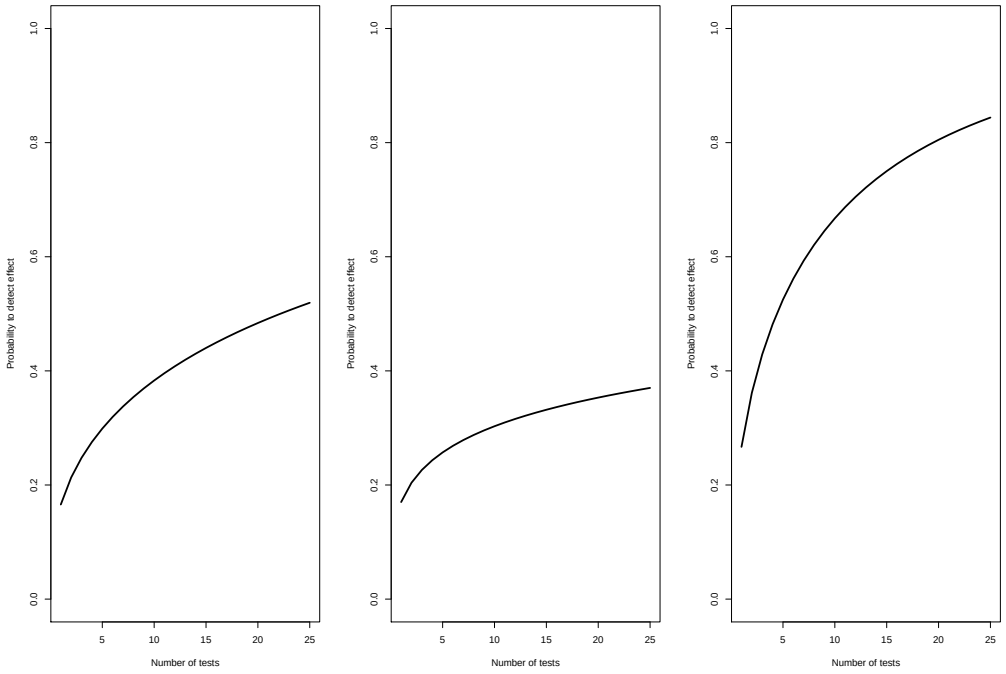

FIG 1. *Probability to falsely find the exposure to be associated with at least one of k mutually independent outcomes at the 5% significance level when Bonferroni correction is used. The bias in the exposure effects is normally distributed with mean θ and standard deviation τ, for $\theta = 0, \tau = 0.1$ (left), $\theta = 0.1, \tau = 0$ (middle) and $\theta = 0.1, \tau = 0.1$ (right); the standard error of the exposure effect is 0.1 for all outcomes.*

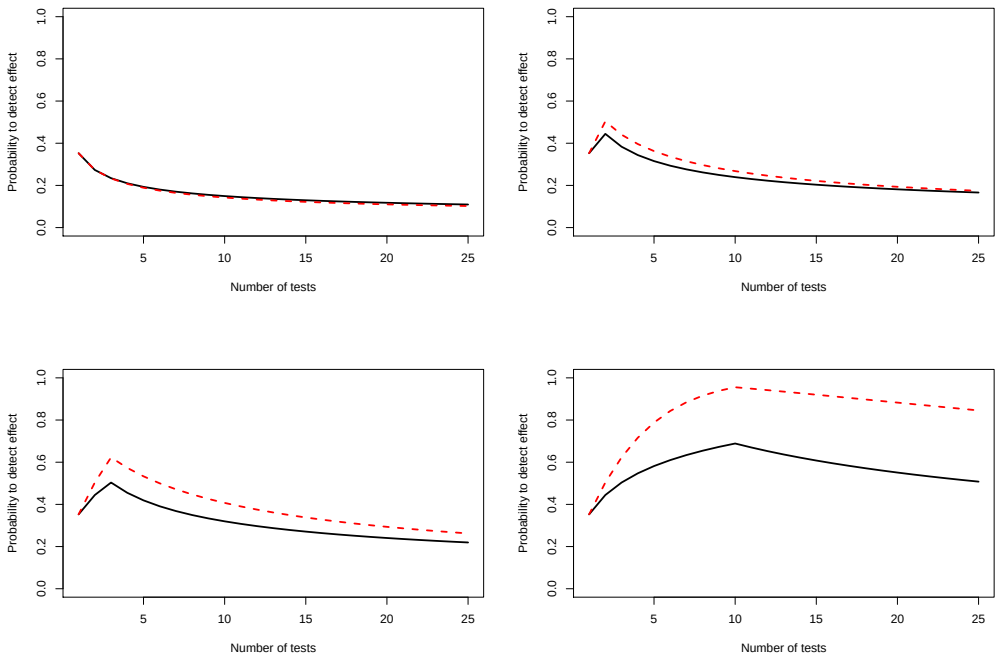

FIG 2. *Probability to find the exposure to be associated with at least one of k mutually independent outcomes (of which the first l are affected by the exposure) at the 5% significance level when Bonferroni correction (solid, black) versus one global test (dashed, red) is used. Upper Left: $l = 1$ outcome is affected by exposure; Upper Right: $l = 2$ outcomes are affected by exposure; Lower Left: $l = 3$ outcomes are affected by exposure; Lower Right: $l = 10$ outcomes are affected by exposure.*