Check for updates

METHOD ARTICLE

Using multiple outcomes in intervention studies for improved trade-off between power and type I errors: the Adjust NVar approach [version 1; peer review: 2 not approved]

Dorothy V. M. Bishop

Department of Experimental Psychology, University of Oxford, Oxford, Oxon, OX2 6GG, UK

V1 First published: 30 Sep 2021, 10:991
<https://doi.org/10.12688/f1000research.73520.1>
Latest published: 13 Jan 2023, 10:991
<https://doi.org/10.12688/f1000research.73520.2>

Abstract

Background

The CONSORT guidelines for clinical trials recommend use of a single primary outcome, to guard against the raised risk of false positive findings when multiple measures are considered. It is, however, possible to include a suite of multiple outcomes in an intervention study, while controlling the familywise error rate, if the criterion for rejecting the null hypothesis specifies that N or more of the outcomes reach an agreed level of statistical significance, where N depends on the total number of outcome measures included in the study, and the correlation between them.

Methods

Simulations were run, using a conventional null-hypothesis significance testing approach with alpha set at .05, to explore the case when between 2 and 12 outcome measures are included to compare two groups, with average correlation between measures ranging from zero to .8, and true effect size ranging from 0 to .7. In step 1, a table is created giving the minimum N significant outcomes (MinNSig) that is required for a given set of outcome measures to control the familywise error rate at 5%. In step 2, data are simulated using MinNSig values for each set of correlated outcomes and the resulting proportion of significant results is computed for different sample sizes, correlations, and effect sizes.

Results

The Adjust NVar approach can achieve a more efficient trade-off between power and type I error rate than use of a single outcome when there are three or more moderately intercorrelated outcome variables.

Conclusions

Where it is feasible to have a suite of moderately correlated outcome measures, then this might be a more efficient approach than reliance on a single primary outcome measure in an intervention study. In effect, it builds in an internal replication to the study. This approach

Open Peer Review

Approval Status

1

2

version 2

(revision)

13 Jan 2023

version 1

30 Sep 2021

view

view

1. **Kristin Sainani** , Stanford University,
Stanford, USA

2. **Daniel Lakens** , Eindhoven University of
Technology, Eindhoven, The Netherlands

Any reports and responses or comments on the article can be found at the end of the article.

can also be used to evaluate published intervention studies.

Keywords

intervention, methodology, statistics, correlated outcomes, power, familywise error rate, multiple comparisons

Corresponding author: Dorothy V. M. Bishop (dorothy.bishop@psy.ox.ac.uk)

Author roles: Bishop DVM: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2021 Bishop DVM. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Bishop DVM. **Using multiple outcomes in intervention studies for improved trade-off between power and type I errors: the Adjust NVar approach [version 1; peer review: 2 not approved]** F1000Research 2021, 10:991 <https://doi.org/10.12688/f1000research.73520.1>

First published: 30 Sep 2021, 10:991 <https://doi.org/10.12688/f1000research.73520.1>

The case against multiple outcomes

The CONSORT guidelines for clinical trials (Moher et al. 2010) are very clear on the importance of having a single primary outcome:

All RCTs assess response variables, or outcomes (end points), for which the groups are compared. Most trials have several outcomes, some of which are of more interest than others. The primary outcome measure is the pre-specified outcome considered to be of greatest importance to relevant stakeholders (such as patients, policy makers, clinicians, funders) and is usually the one used in the sample size calculation. Some trials may have more than one primary outcome. Having several primary outcomes, however, incurs the problems of interpretation associated with multiplicity of analyses and is not recommended.

This advice often creates a dilemma for the researcher: in many situations there are multiple measures that could plausibly be used to index the outcome. A common solution is to apply a Bonferroni correction to the alpha level used to test significance of individual measures, but this is over-conservative if, as is usually the case, the different outcomes are intercorrelated. Alternative methods are to adopt some process of data reduction, such as extracting a principal component from the measures that can be used as the primary outcome, or using a permutation test to derive exact probability of an observed pattern of results. Here I explore a further, very simple, option which I term the “Adjust NVar” approach. The idea is that if one has a suite of outcomes, instead of adjusting the alpha level, one can adjust the number of outcomes that are required to achieve significance at the conventional alpha level of .05 to maintain an overall familywise error rate of 1 in 20 or less.

To illustrate the idea with a realistic example, suppose we are reading a report of a behavioural intervention that is designed to improve language and literacy, and there are 6 measures where we might plausibly expect to see some benefit. The researchers report that none of the outcomes achieve the Bonferroni-adjusted significance criterion of $p < .008$, but two of them reach significance at $p < .05$. Should we dismiss the trial as showing no benefit? We can use the binomial theorem to check the probability of obtaining this result if the null hypothesis is true and the measures are independent: it is 0.033, clearly below the 5% alpha level. But what if the measures are intercorrelated? That is often the case: indeed, it would be very unusual for a set of outcome measures to be independent. A thought experiment helps here. Suppose we had six measures that were intercorrelated at .95 - in effect they would all be measures of the same thing, and so if there was a real effect, most of the measures should show it. Extending this logic in a more graded way, the higher the correlation between the measures, the more measures would need to reach the original significance criterion to maintain the overall significance level below .05.

A simulation script was developed to test these intuitions and to obtain estimates of:

- i) the minimum number of outcome variables in a suite that would maintain the overall familywise error rate at 1 in 20, if each individual measure was evaluated at the significance criterion of .05. This we term MinNSig.
- ii) the power to detect a true effect, if the criterion for rejection the null hypothesis was based on the value of MinNSig identified at step A.

Methods

Correlated variables were simulated using in the R programming language (R Core Team 2020) (R Project for Statistical Computing, RRID:SCR_001905). The script to generate and analyse simulated data is available on <https://github.com/oscci/MinSigVar>. Initially, two approaches to modeling correlated variables were compared, but differences between them proved to be trivial, and so only one is reported here.

Method for simulating outcomes

The *mvnrm* function of the MASS package (Modern Applied Statistics with S, RRID:SCR_019125) was used to generate a set of 12 outcome variables with a specified covariance matrix. For simplicity, all variables were simulated as random normal deviates with SD of 1, and the covariance matrix had a prespecified correlation, r , in all off-diagonal elements. The correlation varied across runs from 0 to .8 in steps of .2, and the number of simulated cases varied from 20 to 110 in steps of 30. Outcomes for Intervention (I) and Control (C) groups differed only in terms of the mean, which was always zero for the group C, and a given effect size, e , for group I. The average observed effect size for all measures in a given condition was computed and used as the basis for comparisons of efficiency between single and multiple measure scenarios.

This method is simple but can lead to unrealistic data: in particular it is possible to have a set of outcomes that are independent of one another ($r = 0$) yet all having the same effect size. In real-world data, one would expect outcomes to be

correlated, especially those that all showed an impact of intervention. Conversely, if a set of outcomes was very highly intercorrelated, then we would expect them all to show a similar intervention effect.

An alternative approach was evaluated to consider such cases, in which the set of 12 outcome measures are simulated as indicators of an underlying latent variable, which mediates the intervention effect. This can be achieved by first simulating a latent variable, with an effect size of either zero, for group C, or e for group I. Observed outcome measures are then simulated as having a specific correlation with the latent variable - i.e. the correlation determines the extent to which the outcomes act as indicators of the latent variable. This can be achieved using the formula:

$$r*L + \sqrt{1-r^2}*E$$

where r is the correlation between latent variable (L) and each outcome, and L is a vector of random normal deviates that is the same for each outcome variable, while E (error) is a vector of random normal deviates that differs for each outcome variable. Note that when outcome variables are generated this way, the mean intercorrelation between them will be r^2 . Thus if we want a set of outcome variables with mean intercorrelation of .4, we need to specify r in the formula above as $\sqrt{.4}$. Furthermore, the effect size for the simulated variables will be lower than for the latent variable: to achieve an effect size, e , for the outcome variables, it is necessary to specify the effect size for the latent variable, e_1 , as e/r . It was found that when this is done, the results with this method were closely similar to those obtained using MASS, for the range of correlations and effect sizes considered here. The exception is for the case where $r = 0$, which is not computable with this method - i.e. it is not possible to have a set of outcomes that are indicators of the same latent factor but which are uncorrelated. As noted above, the case where $r = 0$ is unrealistic in any case, and so for the simulations reported here, the lowest value of r that was included was $r = .2$.

Data reduction

The size of the suite of outcome variables entered into later analysis ranged from 2 to 12. For each suite size, principal components were computed from data from the C and I groups combined, using the base R function *prcomp* from the *stats* package (R Core Team, 2020). Thus, PC2 is a principal component based on the first two outcome measures, PC4 based on the first four outcome measures, and so on.

Power of analyses based on the principal components was compared with power obtained using the Adjust NVar approach, as specified below.

Simulation parameters

10,000 simulations were run for each combination of:

- sample size per group, ranging from 20 to 110 in steps of 30
- correlation between outcome variables, ranging from .2 to .8 in steps of .2
- true effect size, taking values of 0, .3, .5, or .7.

The data generated from each combination of conditions was used to derive results for different sizes of suites of outcome variables, ranging from 2 to 12. Thus, the analysis was first conducted on the first 2 outcome measures, then on the first 3 outcome measures, and so on.

For each set of conditions, on each run, a one-tailed t-test was conducted to obtain a p-value for the comparison between C and I groups, assuming C would be lower. The p-values for outcome measures were rank ordered for each run and each suite size.

Identifying MinNSig

To obtain MinNSig, the results were filtered to include only the runs where the null hypothesis was true, i.e. effect size = 0. Then, the proportion of p-values less than .05 was calculated for each rank for each number of outcome variables, to find the highest rank at which the overall proportion was less than .05. This is the MinNSig.

Table 1 gives a toy example of the logic, using the case where we have either 2 or 4 outcome measures. Columns V1 to V4 show p-values for the t-test comparing the two groups each of the 4 outcome measures. Columns r2.1 and r2.2 show the same p-values rank ordered for just the first two measures; columns r4.1 to r4.4 show the p-values rank ordered for all 4 outcomes. We can then count the number of p-values that are below .05 for all runs (1000 in this case) for each ranked position. With 2 outcomes, if we take just the first ranked (lowest) p-value, the proportion lower than .05 is around .10. For the 2nd ranked p-value, the proportion drops below .05, to .002. Thus, we set MinNSig to 2.

Table 1. Demonstration of how MinNSig is determined. V1 to V4 are p-values from one-sided t-test, one row for each run of simulation in a given condition. Columns with prefix r2 or r4 show the same p-values rank ordered for either the first two columns, or the first four columns. The final two rows show the number and the proportion of values falling below .05 for that column.

run	V1	V2	V3	V4	r2.1	r2.2	r4.1	r4.2	r4.3	r4.4
1	0.877	0.569	0.642	0.661	0.569	0.877	0.569	0.642	0.661	0.877
2	0.272	0.367	0.841	0.954	0.272	0.367	0.272	0.367	0.841	0.954
3	0.993	0.116	0.249	0.414	0.116	0.993	0.116	0.249	0.414	0.993
4	0.366	0.613	0.73	0.401	0.366	0.613	0.366	0.401	0.613	0.73
5	0.735	0.559	0.329	0.715	0.559	0.735	0.329	0.559	0.715	0.735
6	0.781	0.561	0.123	0.052	0.561	0.781	0.052	0.123	0.561	0.781
7	0.805	0.779	0.989	0.991	0.779	0.805	0.779	0.805	0.989	0.991
8	0.97	0.233	0.777	0.151	0.233	0.97	0.151	0.233	0.777	0.97
9	0.129	0.432	0.215	0.005	0.129	0.432	0.005	0.129	0.215	0.432
10	0.897	0.582	0.588	0.64	0.582	0.897	0.582	0.588	0.64	0.897
...	...	...	...	...	...	...	...	...	...	...
999	0.034	0.335	0.238	0.208	0.034	0.335	0.034	0.208	0.238	0.335
1000	0.045	0.086	0.176	0.001	0.045	0.086	0.001	0.045	0.086	0.176
N < .05	.	.	.	.	100	2	185	14	0	0
p < .05	.	.	.	.	0.1	0.002	0.185	0.014	0	0

We can then turn to the case where we have four outcomes: the proportion of the 1st ranked p-values below .05 is .185; the proportion of the second ranked below .05 is .014. Thus again, we set MinNSig to 2. As noted above, when the correlation between variables is zero, we can use the binomial theorem to compute values in the final row; however, when variables are intercorrelated, more p-values will be below .05, and so MinNSig may be higher.

Because MinNSig moves in quantum steps, the effective familywise error rate is often lower than .05. For instance, in the example above with a suite of four outcome measures, MinNSig is set to 2, but this gives $p = .014$, rather than .05.

Computing power using Adjust NVar

For each run of the simulation, and each number of outcome measures, we take the value of MinNSig from the previous step and compute the proportion of p-values below .05, depending on the effect size, sample size and correlation between measures. For effect sizes above zero, this proportion corresponds to the statistical power.

Power using Adjust NVar can be compared to:

- power obtained with a single outcome measure for the same effect size and sample size
- power obtained by using the principal component extracted for this set of outcome measures

Table 2. Values of MinNSig for different suite sizes of outcomes. Entries in body of table show smallest N variables reaching $p < .05$ that preserve familywise error rate at .05 or less. N prefix denotes suite size for a set of outcomes. Corr indicates correlation between outcomes.

corr	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11	N12
0.0	2	2	2	2	2	2	3	3	3	3	3
0.2	2	2	2	2	3	3	3	3	3	3	4
0.4	2	2	2	3	3	3	3	3	4	4	4
0.6	2	2	2	3	3	3	4	4	4	5	5
0.8	2	2	3	3	3	4	4	4	5	5	6

Results

MinNSig

Table 2 shows results from a simulation of the Adjust NVar approach, with the values in the body of the table showing MinNSig, the minimum number of measures that would maintain the overall familywise error rate at 1 in 20, if each individual measure was evaluated at the significance criterion of .05. Because the t-test statistic used to determine p-values is adjusted for sample size, these values are independent of numbers of subjects. In principle, researchers could use **Table 2** to specify in their research protocol the minimum number of outcomes that would reach their significance level in order for the null hypothesis to be rejected.

Power of Adjust NVar approach

Full tables of results for all combinations of parameters are provided in Extended Data. **Figures 1 to 3** plot power vs familywise error rate for different sizes of suite of outcome measures.

Figure 1. Power × Familywise error rate using Adjust NVar method, for small effect size. Symbols denote sample size per group, and colours denote correlation between outcomes (see Key). Vertical dotted lines show power for single outcome at different sample sizes. Horizontal line shows type I error rate of .05.

Figure 2. Power × Familywise error rate using Adjust NVar method, for medium effect size. Symbols denote sample size per group, and colours denote correlation between outcomes (see Key). Vertical dotted lines show power for single outcome at different sample sizes. Horizontal line shows type I error rate of .05.

For these plots, we see power for small, medium and large effect sizes (corresponding to Cohen's d of .3, .5 and .7). An efficient method is one that gives power of .8 or above, and a familywise error rate of .05 or less, i.e. the results should cluster in the bottom right quadrant. Power, which depends on sample size, is shown for a study with a single outcome in the vertical dotted lines, with an alpha of .05 shown in the horizontal dotted line. We can compare by eye how well Adjust Nmin with multiple outcomes compares with a single outcome for the same sample size. With just two outcome measures we obtain a very low familywise error rate but power is generally worse than for the single outcome case, except when the effect size is large. This is because when using Adjust NMin with two outcomes, both outcomes have to achieve $p < .05$. With three outcomes, again Adjust NMin requires we have at least two individual outcomes with $p < .05$: this gives power equivalent to that of a single variable, but a lower familywise error rate is achieved. When the number of outcomes is four or more, the benefit of Adjust NMin over a single outcome becomes more evident, with higher power coupled with a lower familywise error rate. The specific results depend also on the intercorrelation between outcomes (which in turn influences the MinNSig value, see Table 2): a moderate level of intercorrelation (between .4 and .6) generally gives an efficient measure.

Figure 3. Power × Familywise error rate using Adjust NVar method, for large effect size. Symbols denote sample size per group, and colours denote correlation between outcomes (see Key). Vertical dotted lines show power for single outcome at different sample sizes. Horizontal line shows type I error rate of .05.

Figures 4 to 6 give equivalent plots for power from principal components.

The Principal Components plots show all points are to the right of the vertical line denoting power from a single outcome, i.e. this method achieves higher power than a single outcome measure for each size of suite of outcomes. Familywise error rate clusters around .05. If we compare how Adjust NVar compares with Principal Components, with three or more outcomes the power is generally slightly lower, but the tradeoff between power and familywise error (expressed as a ratio) is higher for Adjust NVar.

Discussion

The logic of conventional multiple testing is turned on its head with the Adjust NVar approach, in that instead of adjusting the p-value used for significance (as in the Bonferroni correction, or methods based on False Discovery Rate), we adjust

Figure 4. Power × Familywise error rate using 1st principal component, for small effect size. Symbols denote sample size per group, and colours denote correlation between outcomes (see Key). Vertical dotted lines show power for single outcome at different sample sizes. Horizontal line shows type I error rate of .05.

the number of individual outcome measures that we need to reach the intended significance criterion. This value can be easily computed using the binomial theorem for a given suite size of outcomes if the measures are uncorrelated, but in the context of intervention trials uncorrelated measures is an unrealistic assumption.

One advantage of this approach is that it is more compatible with trials of interventions that are expected to affect a range of related processes, as is common in some fields such as education or speech and language therapy. In such cases, the need to specify a single primary outcome tends to create difficulties, because it is often unclear which of a suite of outcomes is likely to show an effect. Note that the Adjust NVar approach does not give the researcher free rein to engage in p-hacking: the larger the suite of measures included in the study, the higher the value of MinNSig will be. It does, however, remove the need to put all one's eggs in one basket by pre-specifying one measure as the primary outcome.

Figure 5. Power × Familywise error rate using 1st principal component, for medium effect size. Symbols denote sample size per group, and colours denote correlation between outcomes (see Key). Vertical dotted lines show power for single outcome at different sample sizes. Horizontal line shows type I error rate of .05.

A second advantage is that in effect, by including multiple outcome measures, one can improve the efficiency of a study, in terms of the trade-off between power and familywise errors. A set of outcome measures may be regarded as imperfect proxy indicators of an underlying latent construct, so we are in effect building in a degree of within-study replication if we require that more than one measure shows the same effect in the same direction before we reject the null hypothesis.

The comparison with power and familywise error rate from principal components shows that the latter approach is more consistent in improving power over a study with a single outcome than the Adjust NVar approach, regardless of the size of the suite of outcomes, but it does not influence familywise error rate. Variation in familywise error rate is a consequence of the quantum nature of the adjustment with Adjust NVar, where the same value of MinNVar may be used with varying sizes of outcome suite, which can lead to values of familywise error rate well below .05. For instance, obtaining two p-values below .05 in a suite of two outcomes is a more unusual circumstance than obtaining two values this extreme in a

Figure 6. Power × Familywise error rate using 1st principal component, for large effect size. Symbols denote sample size per group, and colours denote correlation between outcomes (see Key). Vertical dotted lines show power for single outcome at different sample sizes. Horizontal line shows type I error rate of .05.

suite of three or four outcomes. Nevertheless, the ratio of power to familywise error is generally higher for Adjust NVar than for principal components.

A possible disadvantage of Adjust NVar over principal components is that this approach is likely to tempt researchers to interpret specific outcomes that fall below the .05 threshold as meaningful. They may be, of course, but this simulation demonstrates that when we create a suite of outcomes that differ only by chance, it is common for only a subset of them to reach the significance criterion. Any recommendation to use Adjust NVar should be accompanied by a warning that a suite of outcomes should be selected as representative of the underlying construct the intervention is designed to influence, in effect serving as replicate measures, all of which should be equally promising as indicators of an intervention effect. If a subset of outcomes show an effect of intervention, this could be due to chance. It would be necessary to run a replication to have confidence in a particular pattern of results.

It is also worth noting that results obtained with this approach depend crucially on assumptions embodied in the simulation that is used to derive predictions. Outcome measures simulated here are normally distributed, and uniform in their covariance structure. It would be possible to generate datasets with different underlying covariance structures to be tested in the same way, but that is beyond the scope of this paper.

Perhaps the main advantage of this approach is that once the values of MinNSig have been specified (as in [Table 2](#)), the method is very simple to apply, and could be used in two ways. First, it can be used *a priori* to specify in a protocol the number of outcomes that would need to achieve the conventional .05 level of significance, in order for the intervention to be deemed effective. This assumes that the researcher already has a rough idea of the degree of intercorrelation between outcome measures, but a range somewhere between .4 and .6 is a reasonable assumption for many behavioural studies. Pre-registering a specific level of MinSigN would help guard against a tendency to explore different kinds of correction for multiple hypothesis testing only after viewing the data ([Lazic 2021](#)).

Second, the simplicity of the approach makes it useful for evaluating published studies that report multiple outcomes but may not have been analysed optimally. We started with the example of a study where there were six outcome measures, none of which met the Bonferroni-corrected significance level of $.05/6 = .008$, but two of which met $p < .05$. From [Table 2](#) we can see that to be confident in a true intervention effect, assuming correlated outcomes, then at least three out of six outcomes need to be significant at the .05 level. In this case, therefore, we do not reject the null hypothesis.

In sum, the Adjust NVar method shows how inclusion of multiple outcomes can be a positive strategy in intervention studies and can give stronger statistical evidence than a single outcome, provided that attention is paid to the need for several outcomes to reach a significance threshold.

Data availability statement

Extended data

OSF: Adjust NVar. <https://doi.org/10.17605/OSF.IO/5T4SE>. (Bishop, 2021).

This project contains the following extended data.

- Extended data.docx (Word version of powertab_methodM.csv, showing power for each combination of parameters, with separate subtables for individual variables ([Figures 1-3](#)) and principal components ([Figures 4-6](#)).

Data are available under the terms of the [Creative Commons Zero “No rights reserved” data waiver](#) (CC0 1.0 Public domain dedication).

Software availability

The script to generate and analyse simulated data is available on <https://github.com/oscci/MinSigVar>.

The OSF project contains the following files.

- toybit.csv - corresponds to [Table 1](#)
- MinNSig_methodM.csv - corresponds to [Table 2](#)
- p_1sided_methodM_allN_allES_allcorr_maxn12_nsim10000_nstep1.csv - output of simulation of 10000 runs of Multiple_outcomes.Rmd script
- Extended data.docx - word version of powertab_methodM.csv
- powertab_methodM.csv - csv version of Extended Data table, with individual variables in columns N1-N12, and principal components in columns PC2-PC12

References

Bishop DVM: **Adjust NVar**. 2021, September 24.

[Reference Source](#)

Lazic SE: **Why Multiple Hypothesis Test Corrections Provide Poor Control of False Positives in the Real World**. *arXiv:2108.04752 [q-Bio, Stat]*. 2021; (August).

[Reference Source](#)

Moher D, Hopewell S, Schulz KF, *et al.*: **CONSORT 2010 Explanation and Elaboration: Updated Guidelines for Reporting Parallel Group**

Randomised Trials. *BMJ (Clinical Research Ed.)*. 2010; **340**(March): c869. [Publisher Full Text](#)

R Core Team: *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2020.

[Reference Source](#)

Venables WN, Ripley BD: *Modern Applied Statistics with S*. Fourth ed. New York: Springer; 2002. 0-387-95457-0.

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 10 November 2021

<https://doi.org/10.5256/f1000research.77175.r97181>

© 2021 Lakens D. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Daniel Lakens

Human-Technology Interaction Group, Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, Eindhoven, The Netherlands

The author discusses more optimal ways to control error rates than the Bonferroni correction when researchers use multiple measures in a study that are positively correlated. In these cases the authors proposed to specify the number of variables that should be significant at the default alpha level (e.g., 0.05) to make sure the overall Type 1 error rate does not exceed 0.05.

The difficulty with correcting for multiple comparisons based on the number of variables that are significant is that you want to prevent a situation where a researcher performs a Stroop effect test, a Simon effect test, and a test for precognition, and conclude the hypothesis is supported because 2 out of 3 tests are significant. The solution the author proposes is to only use this approach if 1) the same latent variable is measured in different measures, and 2) the correlation between these variables can not be zero. I believe this is a valid approach. These assumptions are discussed enough in the article, but on page 11 (of the pdf version), one might want to discuss that *whether* different measures actually are 'replicants' could be a matter of debate among peers. One could imagine a situation where researchers in a field who disagree about measures would also disagree about whether different measures are replicants.

The comparison against the Bonferroni correction is one interesting baseline, but there is a large literature on how to correct for multiple comparisons that is also much more efficient than a Bonferroni correction, and which is the more interesting comparison. A strength of the Bonferroni correction is that it makes no assumptions about the variables that are corrected. But if one is willing to make assumptions, most importantly about the correlation between variables, more efficient approaches are available. How does this approach compare against other correction approaches? Although there are many correction approaches (this seems to be a particularly active field in neuroscience, where multiple comparisons when analyzing brain activation are common, and measures are strongly correlated), the following two references provide a starting point.

Fan, J., Han, X., & Gu, W. (2012). Estimating False Discovery Proportion Under Arbitrary Covariance Dependence. *Journal of the American Statistical Association*, 107(499), 1019–1035¹

Yekutieli, D., & Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82(1), 171–196²

There are several things to consider. First, the proposed simulation based approach fixes the alpha level, and selects the number of variables that need to be significant. Numbers of variables are relatively crude, in that we can pick 1, 2, 3, 4, etc, but not 1.23 variables. Alternative approaches in the literature lower the alpha level. A benefit of these approaches is that the alpha level can be set at any value (e.g., 0.0352) to exactly control the Type 1 error rate. The consequences of this are also clear when we look at the figures and compare the principle component approach with the Adjust NVar approach. The familywise error rate for the Adjust NVar approach is often well below 0.05, while it is controlled at 0.05 in the principle component approach (and it is controlled at 0.05 in other approaches in the literature that lower the alpha level). The author discusses this (e.g., page 7 of the pdf version) in some detail, but the author seems slightly biased towards their own approach, stating that “the tradeoff between power and familywise error (expressed as a ratio) is higher for Adjust NVar.” It is not clear this ratio is a fair evaluation of the methods, and the information is difficult to distill from the figures (a Table with Type 1 error rates and Type 2 error rates would be more useful for this). This ratio is not so easy to summarize in a single sentence, I feel. If a design has 99.9% power, lowering the alpha from 0.05 to 0.02 has little effect on power, but if power is 0.8, lowering the alpha has a greater effect. Typically, the evaluation is done on the required sample size – which type of correction would require the smallest sample size? And then it becomes important to take a long more modern corrections for correlated variables as well.

I could imagine other modern corrections are a bit more efficient, but the author makes a good point that the simplicity of the methods allows one to evaluate published studies. This might be a useful application. However, then I would like to urge the author to add an example. It would be good to see how this approach should be applied in a real use case (e.g., read a justification for why the measures are measuring the same latent construct, and how to make an assumption about the correlation) and how to interpret the results, depending on how many tests are significant.

The current simulation is somewhat limited. The idea that all correlations are identical will not be true in practice, and this will complicate the application of the proposed technique. What is the recommendation in practice? Should authors choose the largest correlation or the smallest correlation? Which approach is more or less conservative? What is the effect of differences in standard deviations, and does it matter if measures with low correlations have larger standard deviations? I am not sure all these factors will have a large influence but users will most likely need to know what to do.

It was very nice to see the paper was written in Rmarkdown. I was able to reproduce the results computationally (I did not repeat the simulations). As a minor comment, skipsim was no longer on line 207 but on 222. The files would be clearer if all simulation code was a separate R script that is run to generate the simulation data. The datafiles generated can then be read in at the top of the Rmd file. The ‘skipsim’ workaround does not improve clarity, and I had a hard time figuring out where the code reads in the data. In row 561, for example, the toybit file is written, but if I read it

in, this is not needed. Separating the creation of data, and reading in data, makes this cleaner. I needed to create an 'Images' folder to run the plot code on line 679 – this folder could be uploaded to the github repo perhaps?

To conclude, I believe this current version of the manuscript needs some additional work, which could include a more extensive discussion of other corrections in the literature, an exploration of additional simulations, and a practical example of how to use this approach when analyzing published studies.

References

1. Fan J, Han X, Gu W: Estimating False Discovery Proportion Under Arbitrary Covariance Dependence. *J Am Stat Assoc.* 2012; **107** (499): 1019-1035 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Yekutieli D, Benjamini Y: Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference.* 1999; **82** (1-2): 171-196 [Publisher Full Text](#)

Is the rationale for developing the new method (or application) clearly explained?

Partly

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Applied statistics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 18 Nov 2022

Dorothy Bishop, University of Oxford, Oxford, UK

I thank the reviewer for such a comprehensive and constructive review. The two reviews

complemented one another very nicely, and helped me see how to restructure and revise the paper to make it clearer.

As background, I should explain that this paper grew out of an attempt to write a primer on how to analyse/evaluate published intervention studies for students/professionals in allied health professions/education. My concern was that Bonferroni correction is often too conservative, but explanations of other methods are typically highly complex, so I was looking for a simpler rule of thumb that could be useful in interpreting published studies. I have emphasised this rationale more in the revised paper.

Response to specific points.

1. The difficulty with correcting for multiple comparisons based on the number of variables that are significant is that you want to prevent a situation where a researcher performs a Stroop effect test, a Simon effect test, and a test for precognition, and conclude the hypothesis is supported because 2 out of 3 tests are significant. The solution the author proposes is to only use this approach if 1) the same latent variable is measured in different measures, and 2) the correlation between these variables can not be zero. I believe this is a valid approach. These assumptions are discussed enough in the article, but on page 11 (of the pdf version), one might want to discuss that *whether* different measures actually are 'replicants' could be a matter of debate among peers. One could imagine a situation where researchers in a field who disagree about measures would also disagree about whether different measures are replicants

Response: This is an important point, which is similar to issues raised by reviewer 1, so I have now extended the simulations to take into account situations where outcomes may not all be indicators of one factor.

2. Alternative methods for correction for multiple comparisons.

Response: The reviewer recommends two references with alternative approaches to multiple comparison correction. As I noted above, one goal was to keep things simple: so I am reluctant to do a full comparison of all methods, especially since it is clear that I need to consider different underlying data structures. I like the relative simplicity of the MEff approach recommended by reviewer 1, and so I am now presenting a comparison of Bonferroni, PCA and MEff.

3. I could imagine other modern corrections are a bit more efficient, but the author makes a good point that the simplicity of the methods allows one to evaluate published studies. This might be a useful application. However, then I would like to urge the author to add an example. It would be good to see how this approach should be applied in a real use case (e.g., read a justification for why the measures are measuring the same latent construct, and how to make an assumption about the correlation) and how to interpret the results, depending on how many tests are significant.

Response: Again, this converges nicely with the recommendations of reviewer 1, and I have now restructured the paper to focus more on this aspect and to give a real-life example.

4. The current simulation is somewhat limited. The idea that all correlations are identical will not be true in practice, and this will complicate the application of the proposed technique. What is the recommendation in practice? Should authors choose the largest correlation or the smallest correlation? Which approach is more or less conservative? What is the effect of differences in standard deviations, and does it matter if measures with low correlations have larger standard deviations? I am not sure all these factors will have a large influence but users will most likely need to know what to do.

Response: I now contrast three models with different patterns of intercorrelation between intervention and outcome. Potentially, there is no end to the options to consider, but I found it reassuring that with the MEff approach, the adjusted alpha was similar for different correlated variables, provided the mean off-diagonal correlation was constant.

5. It was very nice to see the paper was written in Rmarkdown. I was able to reproduce the results computationally (I did not repeat the simulations). As a minor comment, skipsim was no longer on line 207 but on 222. The files would be clearer if all simulation code was a separate R script that is run to generate the simulation data. The datafiles generated can then be read in at the top of the Rmd file. The 'skipsim' workaround does not improve clarity, and I had a hard time figuring out where the code reads in the data. In row 561, for example, the toybit file is written, but if I read it in, this is not needed. Separating the creation of data, and reading in data, makes this cleaner. I needed to create an 'Images' folder to run the plot code on line 679 – this folder could be uploaded to the github repo perhaps?

Response: I have now separated the simulation and generation of corresponding figures from the code to write the paper.

Competing Interests: No competing interests were disclosed.

Reviewer Report 12 October 2021

<https://doi.org/10.5256/f1000research.77175.r96192>

© 2021 Sainani K. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Kristin Sainani

Department of Epidemiology and Population Health, Stanford University, Stanford, CA, USA

The paper presents a method for controlling the familywise error rate when testing multiple outcomes: Adjust NVar. Unlike multiple testing adjustments that lower the p-value threshold, the idea behind Adjust NVar is to require a minimum number of p-values (MinNSig) to meet a nominal $p < .05$ threshold.

When outcomes are independent, MinNSig is defined as follows, where M is the number of outcomes:

$X \sim \text{binomial}(M, 0.05)$

MinNSig is the minimum x for which $P(X \geq x) < 0.05$ or, equivalently, $P(X < x) > 0.95$.

For example, if $M=6$, then:

$P(X=0) = .95^6 = 0.735$

$P(X \leq 1) = P(X=0) + P(X=1) = 0.735 + 6 \cdot .95^5 \cdot .05 = .232 + .735 = .967$

Thus, MinNSig=2.

When outcomes are correlated, there is no simple formula for obtaining MinNSig, so the author has used a simulation approach to account for varying correlation structures.

The idea is novel and has several merits. In particular, the approach closely matches what many researchers already do in the published literature. Many published studies apply a p-value cutoff of 0.05 to multiple outcomes without correcting for multiple testing (or designating a primary outcome). I can envision the AdjustNVar approach being a useful heuristic for re-evaluating published studies that used multiple outcomes but failed to account for multiple testing in any way. For example, if an RCT reported 10 outcomes with only 1 significant result ($p < .05$), readers can easily recognize that this is compatible with a chance finding. But what if the trial found 2 significant results or 3? And what if the outcomes are moderately correlated instead of independent? I can envision researchers using a table such as Table 2 of the AdjustNVar paper to make a quick assessment based on the number of outcomes reported and a rough guess at the correlation structure.

However, I am less convinced of the value of AdjustNVar as a formal tool for controlling the familywise error rate in a planned study. At a minimum, further development and a broader set of simulations would be required to support such a recommendation. The current manuscript describes three alternatives to specifying a single primary outcome in an RCT: (1) Bonferroni adjustment, (2) permutation tests, and (3) use of PCA to derive a single composite outcome. But this ignores existing p-value adjustment methods that are less conservative than Bonferroni. For example, the “M-effective” (Meff) approach accomplishes many of the same goals as AdjustNVar (see: Cheverud (2001)¹, Nyholt (2004)², and Derringer (2018)³). In the Meff approach, one adjusts the p-value threshold by dividing by the effective number of outcomes (Meff) rather than the actual number of outcomes (M). Meff is based on the eigenvalues of the correlation matrix of the outcomes. Where Eigen is the observed vector of eigenvalues from the correlation matrix of the outcomes, Meff is calculated as:

$$\text{Meff} = 1 + (M-1) \cdot (1 - (\text{Var}(\text{Eigen})/M))$$

Bonferroni threshold = α/M

Meff threshold = α/Meff

Like AdjustNVar, Meff is simple and accounts for correlated outcomes. But I believe it has several advantages over AdjustNVar: (1) Meff precisely controls the Type I error rate, whereas AdjustNVar has varying Type I error rates that cannot be precisely controlled by the investigator; (2) Meff accounts for the correlation structure observed in the data, whereas AdjustNVar requires the investigator to guess at the correlation structure; if this guess is far off (which could easily be the case), this would lead to poor Type I error control.

This paper has numerous strengths, including the novelty of the idea; the potential use as a heuristic for re-interpreting flawed published papers; the concision of the writing; and the availability of all code and data. The major limitations of the paper are: (1) it presents an overly narrow set of simulations that do not capture most realistic situations, but then makes overly broad claims based on these simulations. (2) it does not compare AdjustNVar to existing approaches that are less conservative than Bonferroni, such as Meff and (3) it does not address the different reasons why researchers may be including multiple outcomes, but these different reasons lead to markedly different correlation structures.

Specific comments:

1. The paper would benefit from further consideration of the reasons why researchers may include multiple outcomes. The paper focuses on the case where an intervention is “expected to affect a range of related processes.” The simulations make assumptions that match this case, assuming equal correlations across outcomes and equal true effects for each outcome. But researchers may include multiple outcomes for many other reasons, such as: (a) they aren’t sure which process the intervention will affect, (b) they believe the intervention may affect two different processes but they measure each process with several different measurements to “hedge their bets”, or (c) they include a “soft” endpoint in addition to a “hard” endpoint because the “hard” endpoint may occur too rarely. Each of these cases corresponds to different assumptions for the simulations. For example, (b) would be expected to have two clusters of highly correlated variables that are only weakly correlated with each, which will affect MinNSig.
2. The paper suggests that AdjustNVar could be used in study planning—researchers would guess at the correlation structure and set a MinNSig ahead of time. But if they guess the correlation structure incorrectly, such as underestimating the true correlation, then they may choose a MinNSig that does not adequately control Type I error.
3. The “quantum nature” of AdjustNVar is not a desirable characteristic. The researcher is unable to precisely control the Type I error rate. In planning a study in which the correlation is expected to be 0.4, for example, Table 2 would suggest that the researcher should then always choose 9 outcomes over 5-8 outcomes, since 9 maximizes the chances of getting at least 3 p-values $<.05$. This is one reason I prefer Meff, which precisely controls the Type I error rate.
4. This description is misleading: “Should we dismiss the trial as showing no benefit? We can use the binomial theorem to check the probability of obtaining this result if the null hypothesis is true and the measures are independent: it is 0.033, clearly below the 5% alpha level.” The description gives the misleading impression that one would be justified in re-evaluating a paper that used a Bonferroni correction by instead applying the criterion of at least two p-values $<.05$. But doing so would inflate the Type I error rate. Results would have been declared significant if EITHER at least one p-value met the Bonferroni threshold OR at least two p-values were $<.05$ — leading to an effective Type I error rate of 7% (assuming independent outcomes). Note that this example reappears in the discussion and also mistakenly implies that had three p-values been $<.05$, we would have been able to reject the null hypothesis. But this is not the case because the results were already subjected to Bonferroni, and additionally subjecting them to AdjustNVar makes the effective Type I error

rate higher than 5%. AdjustNVar should be applied only to re-evaluate studies that failed to incorporate any adjustments for multiple testing originally.

5. I found Table 1 confusing on first read, and I would recommend simplifying it by focusing on a single number of outcomes rather than both 2 outcomes and 4 outcomes and by removing discussions of ranking p-values. (The p-value ranking isn't important — this is just part of the mechanics of how the algorithm is calculating MinNSig, so I don't think it's needed.) For example, you could just focus on calculating MinNSig for 6 outcomes. Show 6 columns of p-values for the 6 outcomes, and then show a single final column that tabulates the number of p-values $< .05$ for each simulated trial. Then show a frequency table of how many simulations out of 1000 resulted in 0 p-values $< .05$, 1 p-value $< .05$, 2 p-values $< .05$, etc. Then indicate that MinNSig occurs at one number above when the cumulative frequency crosses 95%.
6. I'm unclear as to why the paper focuses on one-tailed tests, which are less common in the literature. I think it would be more useful to present two-tailed tests in Table 2 or to present two tables — one for one-tailed tests and one for two-tailed tests. This makes a difference in a few MinNSig values.
7. Figures 1-3: These figures compare AdjustNVar to a single study outcome. I think the logic behind this comparison is flawed, however. It is comparing apples to oranges. The simulation assumes that, when applying AdjustNVar, ALL variables studied have a true effect. This, in effect, stacks the deck on statistical power for *any* method that considers multiple outcomes rather than a single outcome. For example, I ran a simulation comparing Bonferroni with 6 outcomes compared to a single outcome with $n=50$ per group. When the correlation is < 0.8 , Bonferroni also has more power than the single outcome. And, when I compared the Meff strategy to a single outcome in a variety of scenarios, Meff was always more powerful than a single outcome. I think a more useful comparison would be to directly compare different methods that handle multiple outcomes (e.g., PC to AdjustNVar to Meff).
8. Related to comment (7), I don't think the paper is justified in making this broad claim: "The Adjust NVar approach can achieve a more efficient trade-off between power and type I error rate than use of a single outcome when there are three or more moderately intercorrelated outcome variables." This conclusion is true only when the intervention truly affects ALL outcomes, which is a narrow and arguably unrealistic case. A more realistic scenario is where the intervention works only on a subset of outcomes. In this case, the single variable strategy will be more statistically powerful than the multiple-variable strategies if you choose the right variable.
9. Figures 4-6. Same issue as for Figures 1-3: comparing the principal components composite variable strategy (PC) to a single outcome is flawed because the simulation "stacks the deck" for PC by assuming that all outcomes have a true effect. I believe that Figures 1-6 should focus instead on comparisons of different methods for handling multiple outcomes.
10. The article claims that power is only "slightly lower" for AdjustNVar compared with the PC strategy. However, when I run simulations comparing PC to AdjustNVar, I get consistently higher statistical power for PC and I would characterize the difference as more than just

“slight”. For example, with $N=50$, global $\text{corr}=0.6$, global $\text{ES}=0.3$, and 6 outcomes (two-tailed test), I get power of 35% for AdjustNVar versus 37% for Meff versus 46% for PC.

11. PC is always more powerful than AdjustNVar and Meff when we assume that all outcomes have a true effect. However, PC is not more powerful when we assume that only a subset of outcomes have true effects. For example, if I tweak the above simulation so that only three outcomes out of six have true effects of 0.3, this changes the power to 15% for PC and AdjustNVar, and 28% for Meff. This illustrates why one narrow simulation is insufficient for making general conclusions about the tradeoffs in performance between the different methods.
12. In my simulations, Meff consistently has higher statistical power than AdjustNVar, which is a function of the fact that Meff always has a 5% Type I error rate whereas that of AdjustNVar is variable and sometimes lower than 5%. I don't view it as a strength that AdjustNVar results in arbitrarily lower Type I error rates. It is better for the investigator to be able to precisely control the tradeoff between Type I and Type II error. Meff allows this whereas AdjustNVar does not.
13. Figures 1-6: I found these graphs hard to read as they don't have a clear take-home point. I would suggest removing the single outcome comparisons altogether and then reformatting so that the graphs directly compare AdjustNVar to Meff and PC (three lines). In one graph, hold effect size, sample size, and N outcomes constant, and then show the power of the three methods as a function of increasing correlation. In another graph, hold correlation, sample size, and effect size constant, and show the power of the three methods as a function of N outcomes. And so on. All methods aim to control the familywise error rate at 5%, so this should not be a variable. The fact that AdjustNVar sometimes results in a lower Type I error rate is incidental — it arises as a quirk of the method not as the intent of the researcher.
14. Where AdjustNVar may have an advantage over a method like Meff is for re-evaluating flawed published studies where researchers used multiple outcomes but did not account in any way for multiple testing. An AdjustNVar table such as Table 2 would give readers a quick way to reevaluate such studies without the need for any calculations or access to raw data. I would focus the paper more on this application.

References

1. Cheverud JM: A simple correction for multiple comparisons in interval mapping genome scans. *Heredity (Edinb)*. 2001; **87** (Pt 1): 52-8 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Nyholt DR: A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet*. 2004; **74** (4): 765-9 [PubMed Abstract](#) | [Publisher Full Text](#)
3. Derringer J: A simple correction for non-independent tests. 2018. [Publisher Full Text](#)

Is the rationale for developing the new method (or application) clearly explained?

Partly

Is the description of the method technically sound?

Partly

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Statistics, Sports Medicine

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 18 Nov 2022

Dorothy Bishop, University of Oxford, Oxford, UK

Thanks for the very useful evaluation of this paper, which has prompted further thoughts and a revised manuscript. There was some convergence of views of the two reviewers, although the specific recommendations varied. I thank the reviewer for engaging so thoroughly with the manuscript and for helping improve it.

General point A: AdjustNVar as a heuristic.

Reviewer 1 writes: The idea is novel and has several merits. In particular, the approach closely matches what many researchers already do in the published literature. Many published studies apply a p-value cutoff of 0.05 to multiple outcomes without correcting for multiple testing (or designating a primary outcome). I can envision the AdjustNVar approach being a useful heuristic for re-evaluating published studies that used multiple outcomes but failed to account for multiple testing in any way. For example, if an RCT reported 10 outcomes with only 1 significant result ($p < .05$), readers can easily recognize that this is compatible with a chance finding. But what if the trial found 2 significant results or 3? And what if the outcomes are moderately correlated instead of independent? I can envision researchers using a table such as Table 2 of the AdjustNVar paper to make a quick assessment based on the number of outcomes reported and a rough guess at the correlation structure.

Response: After exploring Meff, I decided not to proceed with the AdjustNVar approach, as I think

a lookup table derived from MEff would achieve the same effect but without the problems arising from the quantal nature of N variables.

General Point B: Need to consider MEff

Reviewer 1 notes that the MEff approach accomplishes many of the same goals as AdjustNVar, and is preferable in many respects.

Response: I had been unaware of MEff, and having consulted the references, I agree it is an elegant solution to the problem I was trying to tackle that avoids some of the limitations of AdjustNVar, particularly the need to assume a given correlation structure. I initially thought this approach had only been used in genetics and not been applied in psychology, but the final reference pointed to the work of Derringer, who has done a great job in a preprint that provides a tutorial in its use. I don't think it is much used in intervention research, which is the context I am particularly interested in, and so I think it is worthwhile updating the article so that it serves as a basic introduction to MEff, with discussion of factors affecting power.

Accordingly, I have changed the focus to compare different methods with a focus on MEff.

Specific comments:

1. The paper would benefit from further consideration of the reasons why researchers may include multiple outcomes. The paper focuses on the case where an intervention is "expected to affect a range of related processes." The simulations make assumptions that match this case, assuming equal correlations across outcomes and equal true effects for each outcome. But researchers may include multiple outcomes for many other reasons, such as: (a) they aren't sure which process the intervention will affect, (b) they believe the intervention may affect two different processes but they measure each process with several different measurements to "hedge their bets", or (c) they include a "soft" endpoint in addition to a "hard" endpoint because the "hard" endpoint may occur too rarely. Each of these cases corresponds to different assumptions for the simulations. For example, (b) would be expected to have two clusters of highly correlated variables that are only weakly correlated with each, which will affect MinNSig.

Response: This really got me thinking and I have now incorporated some further simulations where correlations are not uniform, as also added more discussion of this issue

2. The paper suggests that AdjustNVar could be used in study planning—researchers would guess at the correlation structure and set a MinNSig ahead of time. But if they guess the correlation structure incorrectly, such as underestimating the true correlation, then they may choose a MinNSig that does not adequately control Type I error.

Response: Agreed. This is now dropped.

3. The "quantum nature" of AdjustNVar is not a desirable characteristic. The researcher is unable to precisely control the Type I error rate. In planning a study in which the correlation is expected to be 0.4, for example, Table 2 would suggest that the researcher should then

always choose 9 outcomes over 5-8 outcomes, since 9 maximizes the chances of getting at least 3 p-values $< .05$. This is one reason I prefer Meff, which precisely controls the Type I error rate.

Response: Agreed. AdjustNVar now dropped

4. This description is misleading: "Should we dismiss the trial as showing no benefit? We can use the binomial theorem to check the probability of obtaining this result if the null hypothesis is true and the measures are independent: it is 0.033, clearly below the 5% alpha level." The description gives the misleading impression that one would be justified in reevaluating a paper that used a Bonferroni correction by instead applying the criterion of at least two p-values $< .05$. But doing so would inflate the Type I error rate. Results would have been declared significant if EITHER at least one p-value met the Bonferroni threshold OR at least two p-values were $< .05$ — leading to an effective Type I error rate of 7% (assuming independent outcomes). Note that this example reappears in the discussion and also mistakenly implies that had three p-values been $< .05$, we would have been able to reject the null hypothesis. But this is not the case because the results were already subjected to Bonferroni, and additionally subjecting them to AdjustNVar makes the effective Type I error rate higher than 5%. AdjustNVar should be applied only to re-evaluate studies that failed to incorporate any adjustments for multiple testing originally.

Response: thanks for this clarification. As AdjustNVar is now omitted, this no longer applies.

5. I found Table 1 confusing on first read, and I would recommend simplifying it by focusing on a single number of outcomes rather than both 2 outcomes and 4 outcomes and by removing discussions of ranking p-values. (The p-value ranking isn't important — this is just part of the mechanics of how the algorithm is calculating MinNSig, so I don't think it's needed.) For example, you could just focus on calculating MinNSig for 6 outcomes. Show 6 columns of p-values for the 6 outcomes, and then show a single final column that tabulates the number of p-values $< .05$ for each simulated trial. Then show a frequency table of how many simulations out of 1000 resulted in 0 p-values $< .05$, 1 p-value $< .05$, 2 p-values $< .05$, etc. Then indicate that MinNSig occurs at one number above when the cumulative frequency crosses 95%.

Response: this no longer applies as tables are redone

6. I'm unclear as to why the paper focuses on one-tailed tests, which are less common in the literature. I think it would be more useful to present two-tailed tests in Table 2 or to present two tables — one for one-tailed tests and one for two-tailed tests. This makes a difference in a few MinNSig values.

Response: One-tailed tests have a bad reputation because so often they are misused to just require a lower level of significance when there are no directional predictions, but there are contexts in which they are entirely appropriate, and that includes the kinds of intervention study that is the focus of attention here. You can reasonably predict that an intervention will improve performance rather than worsen it. Reviewer 2 wrote a blogpost about this which I find

convincing: <http://daniellakens.blogspot.com/2016/03/one-sided-tests-efficient-and-underused.html>. I have now explained this further.

7. Figures 1-3: These figures compare AdjustNVar to a single study outcome. I think the logic behind this comparison is flawed, however. It is comparing apples to oranges. The simulation assumes that, when applying AdjustNVar, ALL variables studied have a true effect. This, in effect, stacks the deck on statistical power for *any* method that considers multiple outcomes rather than a single outcome. For example, I ran a simulation comparing Bonferroni with 6 outcomes compared to a single outcome with $n=50$ per group. When the correlation is <0.8 , Bonferroni also has more power than the single outcome. And, when I compared the Meff strategy to a single outcome in a variety of scenarios, Meff was always more powerful than a single outcome. I think a more useful comparison would be to directly compare different methods that handle multiple outcomes (e.g., PC to AdjustNVar to Meff).

Response: Thanks again for helping clarify what is being simulated here. I hope this is clearer in the current article. The figures have been redrawn and I hope are now clearer. Perhaps one takeaway point is that, in being concerned to control type I error, the CONSORT recommendations appear to ignore the gains in power that can be achieved with multiple outcomes.

8. Related to comment (7), I don't think the paper is justified in making this broad claim: "The Adjust NVar approach can achieve a more efficient trade-off between power and type I error rate than use of a single outcome when there are three or more moderately intercorrelated outcome variables." This conclusion is true only when the intervention truly affects ALL outcomes, which is a narrow and arguably unrealistic case. A more realistic scenario is where the intervention works only on a subset of outcomes. In this case, the single variable strategy will be more statistically powerful than the multiple-variable strategies if you choose the right variable.

Figures 4-6. Same issue as for Figures 1-3: comparing the principal components composite variable strategy (PC) to a single outcome is flawed because the simulation "stacks the deck" for PC by assuming that all outcomes have a true effect. I believe that Figures 1-6 should focus instead on comparisons of different methods for handling multiple outcomes.

Response: same as for point 6; this is clearly a key issue, and I think it is clearer now that the alternative models (L2 and L2x) are included.

9. The article claims that power is only "slightly lower" for AdjustNVar compared with the PC strategy. However, when I run simulations comparing PC to AdjustNVar, I get consistently higher statistical power for PC and I would characterize the difference as more than just "slight". For example, with $N=50$, global $\text{corr}=0.6$, global $\text{ES}=0.3$, and 6 outcomes (two-tailed test), I get power of 35% for AdjustNVar versus 37% for Meff versus 46% for PC.

Response: thanks for pushing back on this – this is fair comment.

10. PC is always more powerful than AdjustNVar and Meff when we assume that all outcomes have a true effect. However, PC is not more powerful when we assume that only a subset of outcomes have true effects. For example, if I tweak the above simulation so that only three outcomes out of six have true effects of 0.3, this changes the power to 15% for PC and AdjustNVar, and 28% for Meff. This illustrates why one narrow simulation is insufficient for making general conclusions about the tradeoffs in performance between the different methods.

Response: agreed.

11. In my simulations, Meff consistently has higher statistical power than AdjustNVar, which is a function of the fact that Meff always has a 5% Type I error rate whereas that of AdjustNVar is variable and sometimes lower than 5%. I don't view it as a strength that AdjustNVar results in arbitrarily lower Type I error rates. It is better for the investigator to be able to precisely control the tradeoff between Type I and Type II error. Meff allows this whereas AdjustNVar does not.

Response: fair comment

12. Figures 1-6: I found these graphs hard to read as they don't have a clear take-home point. I would suggest removing the single outcome comparisons altogether and then reformatting so that the graphs directly compare AdjustNVar to Meff and PC (three lines). In one graph, hold effect size, sample size, and N outcomes constant, and then show the power of the three methods as a function of increasing correlation. In another graph, hold correlation, sample size, and effect size constant, and show the power of the three methods as a function of N outcomes. And so on. All methods aim to control the familywise error rate at 5%, so this should not be a variable. The fact that AdjustNVar sometimes results in a lower Type I error rate is incidental — it arises as a quirk of the method not as the intent of the researcher.

Response: Agreed. Graphs reformatted as suggested and I hope are now much easier to interpret.

13. Where AdjustNVar may have an advantage over a method like Meff is for re-evaluating flawed published studies where researchers used multiple outcomes but did not account in any way for multiple testing. An AdjustNVar table such as Table 2 would give readers a quick way to reevaluate such studies without the need for any calculations or access to raw data. I would focus the paper more on this application.

Response: Agreed. The paper has been revised to make this point.

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research