American Journal of Epidemiology
© The Author(s) 2018. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.
Vol. 188, No. 3

Research Letter

SOME DESIRABLE PROPERTIES OF THE BONFERRONI CORRECTION: IS THE BONFERRONI CORRECTION REALLY SO BAD?

The Bonferroni correction (1) for multiple testing is sometimes criticized as being overly conservative. The correction is indeed conservative, and there are uniformly more powerful approaches that preserve type I error of the global null hypothesis (2) (see Appendix). However, there are considerations concerning the Bonferroni correction which suggest that its use, and reporting by its standard—perhaps accompanied by other metrics as well (3)—may not be entirely undesirable.

First, use of the Bonferroni correction is often motivated by the desire to preserve the type I error of the global null hypothesis of all tested associations' being in fact null. By dividing the nominal significance level of the α test (e.g., $\alpha = 0.05$) by the number of tests, one is guaranteed, within a hypothesis-testing framework, to reject the global null hypothesis of no association no more than the proportion α (e.g., $\alpha = 5\%$) of the time when the global null does in fact hold. While this is often the motivation presented for use of the Bonferroni correction, the correction itself does have a much stronger property. Suppose one were testing some number K of potential associations, and after Bonferroni correction, J associations were rejected at the α/K significance level. The standard property of the Bonferroni correction that is often pointed out is (as above) that no more than 5% of the time will one incorrectly conclude, "There is at least 1 true association." However, with J rejections at an α/K significance level, one can in fact also consider the much stronger conclusion that "there are at least J true associations," and one will draw this conclusion, when it is false, at most 5% of the time. This is because even if there were in fact only $J - 1$ true associations, the probability of rejecting J or more would still be less than $[K - (J - 1)] \times \alpha/K < K \times \alpha/K = \alpha$. The fact that this much stronger statement, like the rejection of the global null hypothesis, also has only a 5% error rate gives the Bonferroni correction a much stronger interpretation when results surpass this more conservative threshold.

Second, while the Bonferroni correction can impose a fairly severe penalty when sample sizes are small or when many tests are being conducted, in settings in which sample sizes are very large (such as many major epidemiologic cohort studies) and when only a moderate number of tests are being carried out, use of the Bonferroni correction will in fact often make relatively little difference in the magnitude of effect sizes that can generally be detected. Suppose one were examining a single exposure and its associations with a number of subsequent outcomes using data from a large cohort study, recently referred to as an outcome-wide epidemiologic study (4).

Consider a recent study setting (5) with sample size $n = 3{,}929$ and with $K = 24$ outcomes, with a mean linear and logistic regression coefficient standard error of 0.031 across the various outcomes. In this context, for an outcome with a standard error of 0.031, an effect estimate above 0.061 would suffice to pass the nominal $\alpha = 0.05$ significance level threshold and an effect size above 0.095 would suffice to pass the Bonferroni-corrected significance level threshold, of $\alpha = 0.05/24 = 0.0021$. There is a relatively modest range of effect sizes, 0.061–0.095, for which the nominal significance level would be passed but the Bonferroni-corrected threshold would not be. If variability of the outcomes were similar but the sample size were $n = 10{,}000$, an effect estimate above 0.038 (e.g., an odds ratio of 1.039) would suffice to pass the nominal $\alpha = 0.05$ significance level and an effect size above 0.060 (e.g., an odds ratio of 1.062) would suffice to pass the Bonferroni-corrected significance level, of $\alpha = 0.05/24 = 0.0021$. Here the range of effect estimates for which the nominal significance threshold is passed but the Bonferroni-corrected one is not is even narrower, and arguably in many cases that effect size range is sufficiently narrow to often not be of much scientific or public health importance (e.g., if the odds ratio is not even 1.062, the effect size may be too small to be of much importance). Thus, with large sample sizes, in many settings, if the effect size estimate is sufficient to surpass the nominal threshold of $\alpha = 0.05$, then it will very often be sufficient to pass the Bonferroni-corrected threshold as well.

Indeed, in the study referred to above, with the actual sample size of $n = 3{,}929$, of the 20 P values that surpassed the nominal $\alpha = 0.05$ significance level, 17 of those also surpassed the Bonferroni-corrected significance level of $\alpha = 0.05/24 = 0.0021$. Moreover, as per the first point above, one could then make the statement "There are at least 17 true associations," and, under repeated sampling, statements similarly constructed by reporting the number of Bonferroni-corrected rejections would be false less than 5% of the time.

Of course, just because the Bonferroni correction does not impose a severe penalty on the range of effect sizes that can be detected in some contexts, such as when the sample size is large and a moderate number of tests are being conducted, it does not follow from this that the penalty will always be negligible. In many settings, and perhaps especially in small to medium-sized randomized trials, the sample sizes are often considerably smaller and the Bonferroni correction may constitute a much greater penalty for the relevant effect sizes that can be detected than is indicated here. This will also especially be the case in settings in which the study has been powered specifically to detect an effect for a primary outcome but in which many other secondary outcomes are examined as well. In other settings, even if the sample size is reasonably large, if an extremely large set of tests is being carried out—as is often the case, for example, with genome-wide association studies—then the Bonferroni correction might also likewise impose an especially severe penalty.

However, again in many epidemiologic contexts, especially with large longitudinal cohort studies, the penalty of the Bonferroni correction in terms of the potential effect sizes required to pass various thresholds is often very small and the added advantage of the strength of the conclusions that can be put forward

might be considerable. These considerations, however, need to be weighed within the context, and in light of the specific importance of avoiding false-negative conclusions (6). One also need not definitively choose between using or not using the Bonferroni correction. Investigators can report the actual *P* values themselves, and can then also indicate the number of tests and what the Bonferroni-corrected threshold would be. This allows the reader to assess evidence as compared with both the conventional nominal threshold and the Bonferroni-corrected threshold. One can also, in some contexts, compare the number of tests that pass various *P*-value thresholds to a 95% confidence interval for the number of such "rejections" that would be expected under the global null at different significance levels α, but while preserving the correlation structure among the outcomes (3). There is nothing magical about the α = 0.05 threshold, and these various approaches can also be employed across a range of significance level thresholds.

Of course, none of these metrics are perfect, and the hypothesis-testing framework is itself subject to many limitations and abuses (7). Moreover, evidence needs to be evaluated within studies in light of various biases that might arise, ideally applying bias analysis (7–10), and also across studies in meta-analyses (10–12), but reporting several of these measures that address multiple testing can help in that task of evidence synthesis and evaluation.

ACKNOWLEDGEMENTS

This research was supported by National Institutes of Health grant R01CA222147.

Conflict of interest: none declared.

REFERENCES

1. Dunn OJ. Multiple comparisons among means. *J Am Stat Assoc*. 1961:56(293):52–64.
2. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat*. 1979;6(2):65–70.
3. Mathur M, VanderWeele TJ. New metrics for multiple testing with correlated outcomes. (Harvard University technical report). 2018. https://osf.io/k9g3b. Accessed November 1, 2018.
4. VanderWeele TJ. Outcome-wide epidemiology. *Epidemiology*. 2017;28(3):399–402.
5. Chen Y, Kubzansky LD, VanderWeele TJ. Parental warmth and flourishing in mid-life. *Soc Sci Med*. 2019;220(1):65–72.
6. Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology*. 1990;1(1):43–46.
7. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams and Wilkins; 2008.
8. Lash TL, Fox MP, Fink AK. *Applying Quantitative Bias Analysis to Epidemiologic Data*. 1st ed. (Statistics for Biology and Health). New York, NY: Springer-Verlag New York; 2009.
9. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. *Ann Intern Med*. 2017; 167(4):268–274.
10. Mathur M, VanderWeele TJ. Sensitivity analysis for unmeasured confounding in meta-analyses. *Open Sci Framework*. 2016. (https://osf.io/2r3gm/). Accessed November 20, 2018.
11. Schmidt FL, Hunter JE. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. 3rd ed. Thousand Oaks, CA: Sage Publications; 2014.
12. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7(3):177–188.
13. Gordon A, Glazko G, Qiu X, et al. Control of the mean number of false discoveries, Bonferroni and stability of multiple testing. *Ann Appl Stat*. 2007;1(1):179–190.
14. Frane AV. Are per-family type I error rates relevant in social and behavioral science? *J Mod Appl Stat Methods*. 2015;14(1):12–23.

APPENDIX

The controlled Holm procedure (2) controls the familywise error rate (FWER) and is uniformly more powerful than the Bonferroni correction. In the text it is noted that when using the Bonferroni correction, with *J* rejections at an α/K significance level, one can in fact also consider the much stronger conclusion that "there are at least *J* true associations," and one will draw this conclusion, when it is false, at most 5% of the time. Such statements are also valid under any other procedure that strongly controls the FWER, including those that are uniformly more powerful than the Bonferroni correction, such as the Holm procedure. It might therefore be tempting to conclude that regardless of whether one wants to make standard statements about the probability of at least 1 false-positive rejection, about the number of true associations as above, or both, the Bonferroni correction is obsolete and should be replaced with better FWER control procedures. However, this characterization is misleading, because the Bonferroni correction in fact offers a more stringent form of error control than do most FWER-control alternatives.

Specifically, the Bonferroni correction controls the per-family error rate (PFER), which is the mean number of false-positive rejections divided by the number of tests (13, 14). To illustrate the distinction, suppose the FWER is controlled via the uniformly more powerful Holm procedure (2). Then there is less than a 5% probability of obtaining at least 1 false-positive rejection, but *if* there is at least 1 false positive, there is no guarantee of how many there are; there could be 1 or 100. In contrast, the Bonferroni procedure guarantees that even if there is at least 1 false-positive rejection, there are still fewer than $K \times \alpha$ in expectation. Frane (14) has argued persuasively that in many scientific contexts, every additional false-positive rejection is detrimental, and thus controlling the actual number of false positives (via the PFER) is at least as important as controlling the presence or absence of any false positives (via the FWER). The Bonferroni correction may therefore be valuable in these contexts, even when one has also used more powerful FWER corrections.

Tyler J. VanderWeele[1] and Maya B. Mathur[1,2] (e-mail: tvanderw@hsph.harvard.edu)
[1] *Department of Epidemiology, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA*
[2] *Quantitative Sciences Unit, Biomedical Informatics Research Division, Department of Medicine, School of Medicine, Stanford University, Palo Alto, CA*

DOI: 10.1093/aje/kwy250; Advance Access publication: November 19, 2018