Comment on the multiple problems of multiplicity

Dear Editor:

Streiner’s article (1) rightly called attention to the topic of multiplicity (multiple testing) and discussed some viable approaches to the problem. However, several statements in the article require clarification, as addressed in this letter.

The all-nulls-true scenario

It was claimed that the Bonferroni procedure “assumes that the null hypothesis is true for all of the tests, and this is unreasonable, most especially after a significant omnibus $F$ test” (1). However, the Bonferroni procedure accommodates any proportion of the null hypotheses being true, not just the “worst case” scenario that all null hypotheses are true (2).

A researcher’s predictions—even all of them—can be wrong. After all, any prediction known to be correct a priori would not be worth testing, or at least would not have to be included in a family of adjusted tests. Thus, the all-nulls-true scenario should not be universally dismissed as “unreasonable.” This principle holds even when the omnibus test is statistically significant, because the same sampling error that causes a type I error in an individual test can cause a type I error in the omnibus test. Moreover, most multiple-testing approaches, including the Bonferroni procedure, do not require omnibus tests in the first place (3). When an omnibus test is desired, in some cases, Bonferroni adjustments can be made somewhat less stringent to compensate for omnibus “protection,” but this approach has limitations (4).

Dependence among tests

It was claimed that the Bonferroni procedure “assumes that all of the tests are independent” (1). On the contrary, the Bonferroni and Holm procedures require no assumptions about dependence (2). In fact, unlike some other procedures (e.g., Šidák and Hochberg), the Bonferroni and Holm procedures are valid even for negatively dependent tests (although negative dependence is not always plausible).

It is true that when the goal is to control the familywise type I error rate (FWER, the probability of at least one type I error occurring per family), positive dependence among tests generally increases the Bonferroni procedure’s “conservatism” relative to multistep procedures. But the reason for that conservatism is rarely acknowledged: unlike most other FWER-controlling methods, the Bonferroni procedure also controls the per-family type I error rate (PFER; the average number of type I errors expected to occur per family). Because FWER control treats multiple simultaneous errors as no worse than one error, PFER control is arguably preferable when each type I error is highly consequential. But even if the PFER is disregarded (as it typically is) and the Bonferroni procedure is therefore deemed too conservative, multiplicity typically should still be addressed by some other method.

Note that correlated tests are not immune to multiplicity effects. Indeed, the claim that “the problem of correcting for multiplicity is eliminated” when there is a small number of correlated endpoints (1) is true only if those endpoints are (implausibly) correlated at 1. The claim that simultaneous significances in correlated outcomes “reinforce each other” and “strengthen confidence in the results” (1) is also misguided; when outcomes are positively correlated, a low $P$ value in one outcome increases the likelihood of a low $P$ value in another outcome, even if the null hypotheses are true.

Confusions regarding the Holm procedure

The Holm procedure controls the FWER, not the positive false discovery rate (pFDR) as claimed by Streiner (1). In fact, in the conventional frequentist sense, the pFDR “cannot be controlled by a test procedure,” as Storey (5) explicitly acknowledged when he defined the pFDR.

The Holm procedure does—albeit inefficiently—control the false discovery rate (FDR), which is similar to the pFDR under certain conditions. But any other FWER-controlling method, including the (much older and better-known) Bonferroni procedure, also controls the FDR (2). Thus, even if one conflates FDR and pFDR—a conflation that Storey (5) cautioned against—there is no apparent justification for the statement that the Holm procedure “can be seen as the first—and now perhaps best known—of the techniques to control the pFDR” (1).

Streiner’s (1) description of the Holm procedure also includes the following misstatement: “If there are $k$ significant tests, then their associated $P$ levels are rank ordered…” On the contrary, the $k$ ordered $P$ values should include all $P$ values in the family, not just the “significant” ones. Other descriptions were erroneous as well, e.g., the Šidák procedure was mischaracterized as a “multistep” method (1, Abstract).

CIs

An important advantage of the Bonferroni procedure is that it can easily produce adjusted CIs (6). Adjusted CIs are available for some other methods also, but not for Simes-based procedures (e.g., Hochberg) (3). Of course, one could report unadjusted CIs, but unadjusted CIs could conflict with the adjusted hypothesis tests.

Conclusions

The Bonferroni procedure, though not the best option in all circumstances, does not require fundamentally unreasonable assumptions. In fact, it requires no assumptions whatsoever, other than the validity of the $P$ values. The Bonferroni procedure also has additional advantages, such as PFER control.

Researchers should be skeptical of heuristics that discourage multiplicity adjustment for broad categories of circumstances, e.g., correlated outcomes, “planned” tests, and regression-based analyses; note that regulatory agencies for clinical trials (7, 8) do
not endorse such exemptions. Researchers should instead consider the context-specific consequences of type I and type II errors in the given situation.

Note that the most extreme arguments against multiplicity adjustment that were cited by Streiner (1) come from unempirical articles that seem to fundamentally misunderstand the topic. For example, Schulz and Grimes (9) depicted Bonferroni adjustments as omnibus tests of the “universal null hypothesis” that do not allow significance to be localized to individual tests—a known misconception (2).

It is true that “the decision regarding whether or not to correct for multiple testing is a philosophical one” (1), and many other decisions are similarly subjective, such as what α level is appropriate, which error rate (e.g., PFER, FWER, or FDR) is most relevant, and which tests to include in the family. But the potential for type I error inflation caused by disregarding multiplicity is a mathematical reality that is not up for debate.

The author declared no conflicts of interest.

Andrew V Frane
From the Department of Psychology, University of California, Los Angeles, Los Angeles, CA (E-mail: avfrane@ucla.edu).

REFERENCES


Reply to AV Frane

Dear Editor:

I thank Frane for his thoughtful comment. He raises a number of important points, but I am afraid the commentary is also based in part on a misreading of the article, and statements of “facts” are, at best, subject to debate. To begin with, he disagrees with the statement that the Bonferroni, Holm, and Hochberg procedures “assume” that the null hypothesis is true for all of the tests” and states that, “the Bonferroni procedure accommodates any proportion of the null hypotheses being true, not just the ‘worst case’ scenario that all null hypotheses are true.” Perneger (1), however, states that “The Bonferroni method is concerned with the general null hypothesis (that all null hypotheses are true simultaneously)” (p. 1236). Similarly, Frane also objects to the statement that the Bonferroni and Holm procedures assume the independence of the tests (there is no argument that the Hochberg approach does require it). Abdi (2), on the other hand, states that the Šidák–Bonferroni correction, \[1 - \left(1 - z_{\text{FWER}}\right)^{1/k}\] (where FWER indicates the familywise type I error rate), is derived assuming the independence of tests. He then says that the Bonferroni correction is a simpler approximation of this, introduced because of the difficulty involved in working with fractional powers. This would therefore imply that the same assumption, the independence of tests, would also apply in the latter case, a position echoed by McDonald (3). Although others [e.g., Castelhada et al. (4) and Frane] would disagree, I do not think this issue is as black and white as Frane suggests.

A misreading of the article is exemplified by the fact that the Šidák–Bonferroni procedure is not described in the article as a multi-step method; this adjective was applied only to the Holm and Hochberg procedures, which Frane states were incorrectly described. Most unhelpfully, he does not elaborate in what way these latter procedures have been misrepresented in the article, making any response to this statement impossible. Furthermore, the statement that all null hypotheses in a family are true is “unreasonable” was made within the context of adjusting for multiplicity after a significant omnibus F test. If the overall test is significant, then, by definition, at least one of the null hypotheses must be false. A further misreading of the article occurs in his first footnote. He is correct that the Holm procedure was not designed with the false discovery rate (FDR) or positive false discovery rate (pFDR) in mind, but I did not say it was; the article explicitly states that although “it was developed before the term FDR was introduced, it can be seen as the first—and now perhaps best known—of the techniques to control the pFDR.” Finally, Frane states that my claim that simultaneous significances in correlated outcomes reinforce each other and strengthen confidence in the results is “misguided,” because “when outcomes are positively correlated, a low P value in one outcome increases the likelihood of a low P value in another outcome, even if the null hypotheses are true.” However, this overlooks the converse problem: If 2 (moderately) correlated outcomes both support rejection of the null, then this should give further credence to the effectiveness of the intervention, not weaken it, as a Bonferroni correction would.

The author declared no conflicts of interest.

David L Streiner
From the Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, Ontario (e-mail: streiner@mcmaster.ca).

REFERENCES