Patrick M.M. Bossuyt
Department of Clinical Epidemiology
Biostatistics and Bioinformatics
Academic Medical Center
University of Amsterdam
PO Box 22700
1100 DE Amsterdam, The Netherlands
*Corresponding author. Tel.: +33 (0)142345570; fax: +33 (0) 143268979.
E-mail address: jeremie.cohen@inserm.fr (J.F. Cohen)

References

[1] Cohen JF, Chalumeau M, Cohen R, Korevaar DA, Khoshnood B, Bossuyt PM. Cochran's Q test was useful to assess heterogeneity in likelihood ratios in studies of diagnostic accuracy. J Clin Epidemiol 2015;68:299–306.
[2] Cochran WG. The combination of estimates from different experiments. Biometrics 1954;10:101–29.
[3] Cochran WG. Some methods for strengthening the common χ^2 tests. Biometrics 1954;10:417–51.
[4] Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21:1539–58.
[5] Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;327:557–60.
[6] Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions version 5.1.0 [updated March 2011]. The Cochrane Collaboration; 2011. Available at www.cochrane-handbook.org.
[7] Agresti A. Categorical data analysis. 2nd ed. New York: John Wiley & Sons; 2002.

http://dx.doi.org/10.1016/j.jclinepi.2015.05.025

False discovery rate control is not always a replacement for Bonferroni adjustment (Letter commenting on: J Clin Epidemiol. 2014;67:850-7)

Glickman et al. [1] proposed that Bonferroni adjustments are "difficult to justify on philosophical grounds" and that *false discovery rate* (FDR) [2] control should generally replace *familywise type I error rate* (FWER) control. However, these assertions appear to be based on misunderstandings, only some of which (due to space constraints) are identified in this letter.

For example, it was claimed that Bonferroni adjustments only address the "universal null hypothesis" and are invalid when evaluating individual tests. That is a known misconception [3] contradicted by the very statisticians (Dunn, Miller, and Simes) cited in the report by Glickman et al. [1]. It was also claimed that controlling the FDR at 10% for a collection of studies would offer "some assurance" that at most 10% of the significant findings were false positives, whereas controlling the FWER at 10% would not offer such assurance. That is impossible because any procedure that controls the FWER at α controls the FDR at $\leq \alpha$ [2,3].

Glickman et al. claimed that in a published table (Bombardier et al. [4]), no P-values were significant after Bonferroni correction. However, as Bombardier et al. explicitly noted, 10 P-values were "significant after adjusting for the 45 comparisons within the table using the Bonferroni correction" [4]. Moreover, by double counting 10 redundant P-values, Glickman et al. substantially inflated both the number of uniquely defined P-values in the table (45, not 55) and the number of P-values that were significant using the Benjamini–Hochberg procedure [2].

It also appears that not all the hypotheses in the report by Bombardier et al. [4] were of equal likelihood. By including tests that are known to produce low P-values, the FDR can be artificially lowered—a practice that has been appropriately called "cheating" [5]. Thus, low but unremarkable P-values (e.g., history-of-depression predicts depression; $P < .001$) can help higher P-values (e.g., type-of-insurance predicts depression; $P = .01$) to become significant.

The term "Bonferroni procedure" was used interchangeably with "Bonferroni-type adjustments" and "study-wide error rate control." However, there are numerous more "powerful" methods of FWER control that may be preferable to Bonferroni in certain situations. Note that Bonferroni controls not only the FDR and the FWER but also the *per-family type I error rate* (i.e., the expected number of type I errors).

Benjamini [6] rightly emphasized "matching error rates to inference needs," meaning that the appropriate method of error control depends on which error rates are relevant in the given context. In that regard, FDR control simply does not perform the same job that is performed by Bonferroni. For instance, in screenings, it is often reasonable to tolerate some false positives to limit false negatives, in which case FDR control may be appropriate. However, FDR control is not adequate when strong conclusions are intended (e.g., in confirmatory trials), especially when hypotheses have heterogeneous likelihoods. Furthermore, Bonferroni can produce confidence intervals for point estimates (which are often essential for interpreting results; see http://biostat.mc.vanderbilt.edu/wiki/Main/ManuscriptChecklist), whereas FDR-controlling procedures cannot. In short, FDR control can be a powerful exploratory tool, but the most powerful tool in the box is not the right tool for every job.

Andrew V. Frane
University of California
Los Angeles, CA, USA
Tel.: 310-429-2873; fax: 310-206-5895.
E-mail address: AVFrane@gmail.com

References

[1] Glickman ME, Rao SR, Schultz MR. False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. J Clin Epidemiol 2014;67:850–7.

[2] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc 1995;57:289–300.
[3] Goeman JJ, Solari A. Multiple hypothesis testing in genomics. Stat Med 2014;33:1946–78.
[4] Bombardier CH, Fann JR, Temkin NR, Esselman PC, Barber J, Dikmen SS. Rates of major depressive disorder and clinical outcomes following traumatic brain injury. JAMA 2010;303:1938–45.
[5] Finner H, Roters M. On the false discovery rate and expected type I errors. Biom J 2001;43:985–1005.
[6] Benjamini Y. Simultaneous and selective inference: current successes and future challenges. Biom J 2010;52:708–21.

http://dx.doi.org/10.1016/j.jclinepi.2015.03.025

Response to letter by Frane: “False discovery rate control is not always a replacement for Bonferroni-type adjustment”

To the Editor:

We thank Mr. Frane for his thoughtful comments on our article [1]. In retrospect, we wish Mr. Frane’s comments were sought in conjunction with the peer review of our article which no doubt would have resulted in a clearer presentation. We address the more substantial comments in this response.

First, the Bonferroni procedure was developed out of the Neyman–Pearson testing framework [2] to address a problem of simultaneous inference. The Bonferroni procedure relies on a probability calculation that conservatively calibrates the maximum probability of rejecting a collection of true null hypotheses; if any one hypothesis is rejected with the Bonferroni adjustment applied, the best we can say is that the collection of null hypotheses is not all true but without identifying which ones [3–5]. Thus, we maintain that multiple test adjustments which interpolate *P*-values to individual tests from an omnibus inference are unprincipled.

Second, we made no such claim in our article that controlling the family-wise type I error rate (FWER) at the α level does not offer α-level assurance of false discovery rate (FDR) control. We agree that the use of the Bonferroni procedures (and other FWER adjustments) is no worse at controlling the FDR than the Benjamini–Hochberg (BH) procedure, although the use of these procedures comes at a price—usually in the form of sacrificing power.

We do agree with Mr. Frane that FDR adjustment procedures such as the BH procedure can be abused. Our experience is that the type of abuse he cites is far less common than the practice with the Bonferroni procedure when researchers choose to divide multiple tests into groups of small sizes before performing an FWER adjustment. As we described in our article, the distribution of *P*-values for true alternative hypotheses for any application of the BH adjustment must be maintained, so that cherry-picking tests with low *P*-values to include in the adjustment is, indeed, cheating.

Finally, we agree that some studies can tolerate some false positives and others where the false positive rate needs to be severely limited, such as in confirmatory trials. In the Neyman–Pearson framework in which tests are declared significant or not, a decision rule is being applied. Formally, this means that a test decision must have an associated loss function, and as Mr. Frane implies some studies require loss functions that more strictly limit false positives than others. In this sense, the use of an FDR adjustment, or any type of *P*-value adjustment, needs to be consistent with the losses assumed under type I and type II errors. The alleged controversy raised is no different than assuming a significance level of 0.05 vs. other unconventional levels.

In summary, we agree with Mr. Frane that FDR adjustments (and interval estimates [6]) are not a panacea to the difficult challenges of test multiplicity. However, few health researchers are aware of FDR control and some of the philosophical and practical advantages over FWER control as a framework for inference in multiple testing.

Mark E. Glickman*
Center for Healthcare Organization and Implementation Research
Bedford VA Medical Center
200 Springs Road (152)
Bedford, MA 01730, USA
Department of Health Policy and Management
Boston University School of Public Health
Boston, MA, USA

Sowmya R. Rao
Center for Healthcare Organization and Implementation Research
200 Springs Road (152)
Bedford VA Medical Center
Bedford, MA 01730, USA
Department of Quantitative Health Sciences
University of Massachusetts Medical School
368 Plantation Street
Worcester, MA 01605, USA

Mark R. Schultz
Center for Healthcare Organization and Implementation Research
200 Springs Road (152)
Bedford VA Medical Center
Bedford, MA 01730, USA

*Corresponding author: Center for Healthcare Organization and Implementation Research, 200 Springs Road (152), Bedford, MA 01730, USA. Tel.: 781-687-2875; fax: 781-687-3106.
E-mail address: mg@bu.edu (M.E. Glickman)

References

[1] Glickman ME, Rao SR, Schultz MR. False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. J Clin Epidemiol 2014;67:850–7.
[2] Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference: part I. Biometrika 1928;20A:175–240.