10-12-2020

Misguided Opposition to Multiplicity Adjustment Remains a Problem

Andrew V. Frane
University of California, Los Angeles, avfrane@gmail.com

Recommended Citation

בס"ד

EMERGING SCHOLAR

Misguided Opposition to Multiplicity Adjustment Remains a Problem

Andrew V. Frane
University of California, Los Angeles
Los Angeles, CA

Fallacious arguments against multiplicity adjustment have been cited with increasing frequency to defend unadjusted tests. These arguments and their enduring impact are discussed in this paper.

Keywords: Multiplicity, multiple comparisons, multiple testing

Introduction

Since Fisher (1935) suggested Bonferroni-type adjustment to account for multiple significance tests, a sophisticated literature has developed on how to address the problem of multiplicity. However, many studies involving multiple tests have been conducted without accounting for multiplicity, a practice advocated in highly influential anti-adjustment literature.

Perneger (1998) and Rothman (1990) argued not only against the Bonferroni procedure but also against the general principle of multiplicity adjustment in a defense of failure to adjust for multiplicity. The number of citations of these articles per year has trended upward over time (see Figure 1). Rubin (2017) cited Perneger (1998) and Rothman (1990) in defense of the claim it is inappropriate to adjust for multiple hypotheses. Glickman et al. (2014) and Nakagawa (2004) acknowledged the utility of controlling the false discovery rate in certain contexts but dismissed the importance of controlling the familywise error rate altogether. False discovery rate control, though appropriate in some circumstances, is not an adequate substitute for familywise error rate control in general (Dmitrienko et al., 2010, p.

39; Finner & Roters, 2001; Frane, 2016; Meijer & Goeman, 2016; Benjamini, 2010; Benjamini & Hochberg, 1995).

      There is a reluctance to apply multiplicity adjustment, because it reduces statistical power and requires either investing in larger samples or settling for lower power. Given the constant pressure to obtain publishable, statistically significant discoveries, perhaps it should not be surprising that anti-adjustment arguments are so popular. Anti-adjustment articles may appeal to naïve researchers because they are written in plain language, appear in non-statistical journals (with rare exceptions; Saville, 1990), and rely on appeals to "common sense" (e.g., Perneger, 1998, p. 1236). Moreover, the recommendations in anti-adjustment articles are
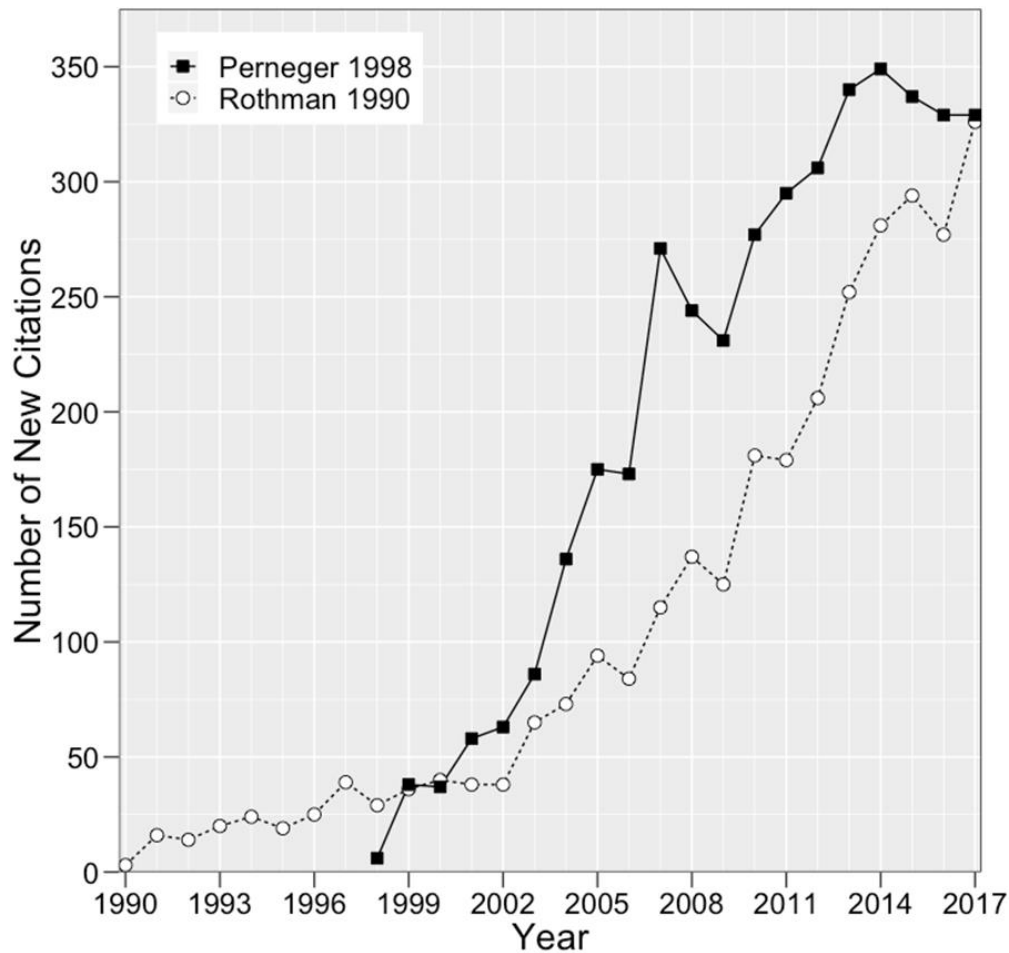
Figure 1. Number of new citations of Perneger (1998) and Rothman (1990) in each year (as per Google Scholar, February 8, 2018)

3

often simple heuristics requiring little thought to implement because they advocate forgoing adjustment altogether in nearly all circumstances, with little to no consideration of contextual factors (such as the goals of the study, how the results will be used to make decisions or draw conclusions, and whether there is a hierarchical structure to the testing).The aim of the present paper is not to establish procrustean rules about how multiplicity should be handled in all contexts. Rather, the aim is to document the prevalence and impact of fallacious arguments against the principle of multiplicity adjustment, and to provide information to counter the use and proliferation of those arguments.

The phrase "multiplicity adjustment" is used in this paper to mean any sound method of addressing multiplicity. This broad definition accommodates methods not involving adjustment, per se, of p-values or nominal alpha levels (e.g., certain sequential testing methods, when the sequence is defined in a pre-registered study protocol). Although multiplicity adjustment is typically discussed in the context of null hypothesis testing, the same principles may apply when using confidence intervals, rather than p-values, as the primary basis for inference or decision-making (Phillips et al., 2013). Thus, contrary to some suggestions (e.g., Huisingh & McGwin, 2012), examining effect-size estimates and confidence intervals rather than only p-values—though generally a good idea—does not in itself eliminate the problem of multiplicity. Many adjustment procedures are applicable to confidence intervals (e.g., Dunn, 1958, 1961; Dunnett, 1955; Tukey, 1953; Westfall, 1985).

Misconceptions Underlying Anti-Adjustment Arguments

Regarding the Universal Null Hypothesis

Some anti-adjustment arguments (e.g., Perneger, 1998; Savitz & Olshan, 1995) included the false claim that Bonferroni-type adjustments only allow inference about the "universal null hypothesis," i.e., about whether the null hypotheses are true for all tests—a view that Goeman and Solari (2014, p. 1955) identified as a myth. For example, Perneger (1998) claimed if two groups are compared on 20 variables and at least one p-value is significant at the Bonferroni-adjusted level, "We can say that the two groups are not equal for all 20 variables, but we cannot say which, or even how many, variables differ…A clinical equivalent would be the case of a doctor who orders 20 different laboratory tests for a patient, only to be told that some are abnormal, without further detail" (p. 1236). That description would be true of a single omnibus test, not of multiple Bonferroni-adjusted tests. Bonferroni adjustments, and many similar methods, allow statements to be made

about each hypothesis because they control the familywise error rate in the strong sense, meaning even if only some of the individual null hypotheses are true (Goeman & Solari, 2014). Classical Bonferroni adjustment also controls the per-family error rate (i.e., the expected number of Type I errors), which is a stricter standard than the familywise error rate (Frane, 2015a).

Rothman (1990) also criticized multiplicity adjustment for allegedly only being relevant to the universal null hypothesis. Rothman suggested that even entertaining a universal null hypothesis would be fundamentally absurd: "Whereas we can imagine individual pairs of variables that may not be related to one another, no empiricist could comfortably presume that randomness underlies the variability of all observations…To entertain the universal null hypothesis is, in effect, to suspend belief in the real world and thereby to question the premises of empiricism" (p. 44-45). However, even if only two null hypotheses are true, the familywise error rate can be inflated to approximately twice the nominal level. Thus, addressing Type I error inflation does not require ascribing all observable associations in the world to pure randomness. Moreover, even if all null hypotheses are false, random variation can still substantially affect observations, as observed associations vary in magnitude—and sometimes direction—from one sample to the next.

There have been numerous citations of Perneger's (1998) and Rothman's (1990) claims about the universal null hypothesis (e.g., Armstrong, 2014; Berry, 2012; De Pablo-Fernandez et al., 2017; Glickman et al., 2014; Jenkins et al., 2009; O'Connor et al., 2009, Ostendorf et al., 2017; Racette et al., 2005; Shulz & Grimes, 2005; Sinclair et al., 2013; van Gils et al., 2009; Zintzaras & Lau, 2008). For instance, Armstrong's (2014) endorsement of Perneger's claim about the universal null hypothesis was in turn cited by several others (e.g., Day & Thorn, 2017; Kim et al., 2015; Ozcan et al., 2017) to defend unadjusted tests.

The claims about the universal null hypothesis by Perneger (1998) and Rothman (1990) have also been cited, without critique, in textbooks (e.g., Ahlbom, 1993, p. 52; Aschengrau & Seage, 2014, pp. 322-323; Shulz & Grimes, 2006, p. 192), and an education research group at Stanford University responded to criticism of their unadjusted testing by claiming adjustment is unnecessary when the universal null hypothesis is not of interest (Center for Research on Education Outcomes, n.d.), citing Perneger and Rothman.

Regarding the Inherent Implausibility of Chance Associations

Many of Rothman's (1990) objections to multiplicity adjustment apparently reflect a more general objection to Type I error control and to any concern that observed

associations in a sample might arise by chance. In Rothman's view, "Being impressed by an extreme result should not be considered a mistake in a universe brimming with interrelated phenomena" (p. 46). It is true that associations are plentiful in the universe, but finite samples can contain misleading associations that do not accurately reflect real effects in the population. If that were not the case, then there would be no need for inferential statistics at all—even in the absence of multiplicity. Yet, Rothman implied that misleading associations are inherently unlikely, at least in biological data.

Rothman (2014) further opined: "If one is studying experiments on psychic phenomena, skepticism about the results might lend support to multiplicity adjustments. If one is studying physiologic effects of pharmaceutical agents, real associations are to be expected and the adjustments are more difficult to defend" (p. 1063). On the contrary, multiplicity adjustment is a mathematical correction based on the number of associations examined, not an expression of skepticism based on the type of associations examined. A single positive test of psychic phenomena would presumably merit considerable skepticism, even if there were no multiplicity to adjust for. Moreover, disregarding multiplicity when evaluating the efficacy of pharmaceutical products would be in direct opposition to the guidelines of regulatory agencies (Committee for Proprietary Medicinal Products, 2002; Food and Drug Administration [FDA], 1998).

There are situations in which it is appropriate to incorporate prior probabilities into the analysis. But that is not achieved by simply ignoring multiplicity. One might argue that "strictly true" null hypotheses (meaning there is no effect even negligibly different from zero in either direction) are in fact rare in biological contexts, and focus should rather be on effect sizes rather than on p-values. But even in that case, multiplicity adjustment would likely be useful for computing simultaneous confidence intervals for the effect sizes.

Regarding Statistical Power and Type II errors

Because the purpose of null hypothesis testing is to protect against spurious discoveries, it would be nonsensical to defend the use of an arbitrarily high alpha level by noting that high alpha levels make discoveries easier to claim. Yet, this argument is frequently advanced, defending inflated familywise alpha levels by noting that unadjusted tests provide more statistical power and lower chance of Type II error. For instance, Fekkes et al. (2006) stated "No adjustment for multiple comparisons, such as the Bonferroni correction, was done, because this would result in an increase in Type II errors, that is, finding a true difference and not

considering this significant (Perneger, 1998)" (p. 1570). Roberts et al. (2011) offered a similar defense of their unadjusted testing: "To avoid Type II errors no adjustment was made for multiple comparisons (Perneger, 1998)" (p. 1558).

Berk, Dean, et al. (2014, p. 360), Berk, Daglas, et al. (2017, p. 415) Carral-Fernández et al. (2016, p. 232), Cotton, Gleeson, et al. (2010, p. 261), Cotton, Lambert, et al. (2013, p. 3), González-Blanch et al. (2015, p. 22), Marion-Veyron et al. (2015, p. 165), Mossaheb et al. (2013, p. 164), and Rajapakse et al. (2014, p. 3) included the following sentence word-for-word: "No adjustments were made for multiple comparisons because they can result in a higher type II rate [sic], reduced power, and increased likelihood of missing important findings (Rothman, 1990)". Nearly identical sentences have appeared in Allott et al. (2015, p. 130), Smyth et al. (2015, p. 889), and others.

Perneger (1998) proposed several scenarios in which multiplicity adjustment would allegedly cause catastrophic Type II errors. Some of those scenarios were nonsensical and bore no resemblance to contexts in which multiplicity adjustment would actually be applied, e.g.: "In a clinical setting, a patient's packed cell volume might be abnormally low, except if the doctor also ordered a platelet count, in which case it could be deemed normal" (p. 1236). Some other scenarios Perneger proposed were more vaguely defined. For example, Perneger warned by applying multiplicity adjustment, "an effective treatment may be deemed no better than placebo" (p. 1236). It is not clear how that would happen, because neither the multiple tests nor the structuring thereof was defined in the scenario. In many cases, testing can be structured so the familywise error rate is controlled without sacrificing statistical power in the primary test of treatment efficacy (Committee for Proprietary Medicinal Products, 2002). In other cases, there is effectively no Type I error inflation to adjust for, because unanimous statistical significance is required on all outcomes simultaneously for the treatment to be approved. In some other cases, multiplicity adjustment is required—and for good reason. For instance, if a treatment is compared to placebo on five outcomes, any one of which on its own could earn approval for the treatment, then without adjustment the probability of erroneously declaring the treatment effective is approximately 23% (given a nominal alpha level of .05, true null hypotheses, roughly independent outcomes, and satisfaction of assumptions).

Noting that looser Type I error control can provide greater statistical power is a trivial and unpersuasive argument for sacrificing statistical rigor. Although statistical power is important, the proper way to limit Type II errors is by using an adequate sample size—not by allowing Type I errors to be arbitrarily inflated (Committee for Proprietary Medicinal Products, 2002; FDA, 1998).

In some early-stage research, it may not be feasible to collect a sample large enough to provide ample statistical power while stringently controlling for multiplicity. But in such cases, rather than ignoring multiplicity to make observed trends appear significant, it would be more appropriate to refrain from making inferential claims until the trends are confirmed in a legitimately higher-powered study. Indeed, contrary to some suggestions (e.g., Aschengrau & Seage, 2014, p. 323; Savitz, 2003, p. 249), statistical nonsignificance does not necessarily imply that the null hypothesis must be accepted per se (in the epistemic sense) without any further investigation. Rather, statistical nonsignificance means the null hypothesis cannot be rejected based on the present evidence. Yet, Rothman (1990) claimed multiplicity adjustment "shields some observed associations from more intensive scrutiny by labeling them as chance findings" (p. 46). Although that claim may accurately depict how some researchers misinterpret or misuse statistical nonsignificance in some cases—whether multiplicity is present or not—it does not constitute a legitimate criticism of the principle of multiplicity adjustment.

Regarding "Arbitrarily" Defined Families

A popular anti-adjustment argument that resembles the fallacy of slippery slope is as follows: The number of tests to adjust for is arbitrary because that family of tests could theoretically be extended to include all the tests conducted in a given researcher's career, or all the tests reported in a given journal (e.g., Feise, 2002; Moran, 2003; Perneger, 1998; Rubin, 2017; Savitz, 2003, pp. 252-253; for similar arguments, see Huisingh & McGwin, 2012; Rothman, 1990). Considering all the tests conducted in an investigator's career or in the history of a journal would indeed be an extreme way to define the family in most cases, and the latter would present the challenge of accounting for publication bias. But considering each test in isolation would be an extreme approach in its own right. For typical applications, a middle ground is likely the most sensible strategy (Miller, 1981, pp. 31-32). The typical consumer of a study containing multiple tests is presumably interested in the results of a particular investigation—not in the results of the author's entire career or of the journal's entire history. That said, if in a particular case there were some compelling reason to interpret results in the context of a researcher's entire career, then it could in fact make sense to adjust inference accordingly. Notwithstanding situations where the definition of the family is dictated by some regulatory agency or other authority, "There are no hard-and-fast rules for where the family lines should be drawn, and the statistician must rely on his [or her] own judgment for the problem at hand" (Miller, 1981, p. 35).

The grouping of tests into families is contextually dependent and somewhat subjective, but not completely arbitrary. Note that the same description— "somewhat subjective, but not completely arbitrary"—could just as easily apply to numerous other a priori decisions, such as what sample size is sufficient, what minimum effect size to consider clinically significant, and what overall alpha level (.05 or some other level) is appropriate. Just as those decisions can be made in a principled way, so can decisions regarding the definition of the family. Contrary to Perneger's (1998) claim "Most proponents of the Bonferroni method would count at least all the statistical tests in a given report as a basis for adjusting P values" (p. 1236), it is doubtful any competent statistician would recommend, for example, adjusting the confirmatory test of primary interest to account for a set of descriptive follow-up tests (Committee for Proprietary Medicinal Products, 2002). How the family should be defined may be debatable in some cases, but that does not mean that any definition of the family is as good as another.

Regarding Planned Tests

It is often said hypothesis tests planned a priori do not require multiplicity adjustment. Statements such as the following, by Fish et al. (2007), are common in the scientific literature: "Whilst it is true that if the Bonferroni adjustment was applied in the following analysis, none of the associations would reach the corrected threshold, there are views strongly opposing the use of such corrections in analyses where a priori hypotheses exist (Perneger 1998)" (p. 1325). Moreover, many authors of textbooks on applied statistics have explicitly recommended not adjusting for multiplicity if the tests were planned (e.g., Ha & Ha, 2012, p. 206; McKillup, 2012, p. 163; Pagano, 2013, p. 422; Rutherford, 2011, p. 76; Scheff, 2016, p. 112). However, there is no apparent scientific basis for that recommendation. For a critique of the "planned-hypotheses exemption from multiplicity adjustment" see Frane (2015b, pp. 6-7).

If no specific tests are planned, then the number of potential tests for the researcher to choose from may be indeterminate, making meaningful adjustment impossible (Hochberg & Tamhane, 1987, p. 10). In that situation, there should not be a false sense of security that Type I error inflation can be prevented merely by adjusting for the tests that were formally conducted.

Conclusion

Although anti-adjustment arguments are frequently cited in scientific literature, they are based largely on misconceptions and, perhaps in some cases, willful misrepresentations. Researchers should be wary of citing an opinion as justification for a particular approach. Educators and textbook authors should warn students about common misconceptions regarding multiplicity. Reviewers and editors should be aware such misconceptions are prevalent in the literature and should combat the propagation of those misconceptions whenever possible. For instance, when reviewing a manuscript, they should be on the lookout for citations of papers that serve as go-to references for researchers seeking to shield their unadjusted testing from criticism (e.g., Preneger, 1998; Rothman, 1990).

Once an anti-adjustment paper has been published, other researchers can write critical letters in response. However, such letters typically receive much less attention than the offending article itself. For example, a letter by Aickin (1999, p. 127) noted that Perneger's (1998) paper "consists almost entirely of errors" and a letter by Bender and Lange (1998) was similarly critical of Preneger's paper—though those letters could not stop its growing influence (as evident from Figure 1).

There is widespread concern in the sciences (e.g., Baker, 2016) that too many findings are not replicable and that there is a high prevalence of Type I errors in the literature. Naturally, neglecting multiplicity exacerbates those problems (as noted by Bretz & Westfall, 2014; Forstmeier et al., 2016; Young, 2009). Therefore, researchers and statisticians have a scientific responsibility to directly confront bad practice and misguided thinking concerning multiplicity.

References

Ahlbom, A. (1993). *Biostatistics for epidemiologists*. Boca Raton, FL: Lewis Publishers.

Aickin, M. (1999). Other method for adjustment of multiple testing exists. *BMJ, 318*, 127-128. doi: 10.1136/bmj.318.7176.127a

Allott, K. A., Rapado-Castro, M., Proffitt, T.-M., Bendall, S., Garner, B., Butselaar, F., Markulev, C., Phassouliotis, C., McGorry, P. D., Wood, S. J., Cotton, S. M., & Phillips, L. J. (2015). The impact of neuropsychological functioning and coping style on perceived stress in individuals with first-episode psychosis and healthy controls. *Psychiatry Research, 226*(1), 128-135. doi: 10.1016/j.psychres.2014.12.032

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optics, 34*(5), 502-508. doi: 10.1111/opo.12131

Aschengrau, A., & Seage, G. R., III. (2014). *Essentials of epidemiology in public health* (3rd edition). Burlington, MA: Jones & Barlett Learning.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature, 533*(7604), 452-454. doi: 10.1038/533452a

Bender, R., & Lange, S. (1998). Rapid response: What's wrong with arguments against multiplicity adjustment (Letter to the editor concerning BMJ 1998;316:1236-1238). *BMJ*. Retrieved from https://www.bmj.com/rapid-response/2011/10/27/whats-wrong-arguments-against-multiplicity-adjustments

Benjamini, Y. (2010). Simultaneous and selective inference: Current successes and future challenges. *Biometrical Journal, 52*(6), 708-721. doi: 10.1002/bimj.200900299

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodology), 57*(1), 289-300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Berk, M., Daglas, R., Dandash, O., Yücel, M., Henry, L., Hallam, K., Macneil, C., Hasty, M., Pantelis, C., Murphy, B. P., Kader, L., Damodaran, S., Wong, M. T. H., Conus, P., Ratheesh, A., McGorry, P. D., & Cotton, S. M. (2017). Quetiapine v. lithium in the maintenance phase following a first episode of mania: Randomised controlled trial. *The British Journal of Psychiatry, 210*(6), 413-421. doi: 10.1192/bjp.bp.116.186833

Berk, M., Dean, O. M., Cotton, S. M., Jeavons, S., Tanious, M., Kohlmann, K., Robbins, J., Cobb, H., Ng, F., Dodd, S., Bush, A. I., & Malhi, G. S. (2014). The efficacy of adjunctive N-acetylcysteine in major depressive disorder: A double-blind, randomized, placebo-controlled trial. *The Journal of Clinical Psychiatry, 75*(6), 628-636. doi: 10.4088/JCP.13m08454

Berry, D. (2012). Multiplicities in cancer research: Ubiquitous and necessary evils. *Journal of the National Cancer Institute, 104*(15), 1125-1133. doi: 10.1093/jnci/djs301

Bretz, F., & Westfall, P. H. (2014). Multiplicity and replicability: Two sides of the same coin. *Pharmaceutical Statistics, 13*(6), 343-344. doi: 10.1002/pst.1648

Carral-Fernández, L., González-Blanch, C., Goddard, E., González-Gómez, J., Benito-González, P., & Bustamante-Cruz, E. (2016). Planning abilities in

patients with anorexia nervosa compared with healthy controls. *The Clinical Neuropsychologist, 30*(2), 228-242. doi: 10.1080/13854046.2016.1147603

Center for Research on Education Outcomes. (n.d.). *CREDO response to critique for multiple comparisons adjustment*. Retrieved from https://web.stanford.edu/group/credo/pdfs/CREDOmethodsresponse.pdf

Committee for Proprietary Medicinal Products. (2002). *Points to consider on multiplicity issues in clinical trials* (CPMP/EWP/908/99). London: European Agency for the Evaluation of Medicinal Products. Retrieved from https://www.ema.europa.eu/en/documents/scientific-guideline/points-consider-multiplicity-issues-clinical-trials_en.pdf

Cotton, S. M., Gleeson, J. F. M., Alvarez-Jimenez, M., & McGorry, P. D. (2010). Quality of life in patients who have remitted from their first episode of psychosis. *Schizophrenia Research, 121*(1-3), 259-265. doi: 10.1016/j.schres.2010.05.027

Cotton, S. M., Lambert, M., Berk, M., Schimmelmann, B. G., Butselaar, F. J., McGorry, P. D., & Conus, P. (2013). Gender differences in first episode psychotic mania. *BMC Psychiatry, 13*(82). doi: 10.1186/1471-244X-13-82

Day, M. A., & Thorn, B. E. (2017). Mindfulness-based cognitive therapy for headache pain: An evaluation of the long-term maintenance of effects. *Complementary Therapies in Medicine, 33*, 94-98. doi: 10.1016/j.ctim.2017.06.009

De Pablo-Fernandez, E., Tur, C., Revesz, T., Lees, A. J., Holton, J. L., & Warner, T. T. (2017). Association of autonomic dysfunction with disease progression and survival in Parkinson disease. *JAMA Neurology, 74*(8), 970-976. doi: 10.1001/jamaneurol.2017.1125

Dmitrienko, A., Bretz, F., Westfall, P. H., Troendle, J., Wiens, B. L., Tamhane, A. C., & Hsu, J. C. (2010). Multiple testing methodology. In A. Dmitrienko, A. C. Tamhane, & F. Bretz (Eds.), *Multiple testing problems in pharmaceutical statistics* (pp. 35-130). Boca Raton, FL: Chapman & Hall.

Dunn, O. J. (1958). Estimation of the means of dependent variables. *Annals of Mathematical Statistics, 29*(4), 1095-1111. doi: 10.1214/aoms/1177706443

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association, 56*(293), 52-64. doi: 10.1080/01621459.1961.10482090

Dunnett. C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association, 50*(272), 1096-1121. doi: 10.1080/01621459.1955.10501294

Feise, R. J. (2002). Do multiple outcome measures require p-value adjustment? *BMC Medical Research Methodology, 2*(8). doi: 10.1186/1471-2288-2-8

Fekkes, M., Pijpers, F. I. M., Fredriks, A. M., Vogels, T., & Verloove-Vanhorick, S. P. (2006). Do bullied children get ill, or do ill children get bullied? A prospective cohort study on the relationship between bullying and health-related symptoms. *Pediatrics, 117*(5), 1568-1574. doi: 10.1542/peds.2005-0187

Finner, H., & Roters, M. (2001). On the false discovery rate and expected Type I errors. *Biometrical Journal, 43*(8), 985-1005. doi: 10.1002/1521-4036(200112)43:8<985::AID-BIMJ985>3.0.CO;2-4

Fish, J., Evans, J. J., Nimmo, M., Martin, E., Kersel, D., Bateman, A., Wilson, B. A., & Manly, T. (2007). Rehabilitation of executive dysfunction following brain injury: "Content-free" cueing improves everyday prospective memory performance. *Neuropsychologia, 45*(6), 1318-1330. doi: 10.1016/j.neuropsychologia.2006.09.015

Fisher, R. A. (1935). *The design of experiments*. Edinburgh, UK: Oliver & Boyd.

Food and Drug Administration. (1998). *Guidance for industry: E9 statistical principles for clinical trials*. Washington, DC: United States Department of Health and Human Services. Retrieved from https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e9-statistical-principles-clinical-trials

Forstmeier, W., Wagenmakers, E.-J., & Parker, T. H. (2016). Detecting and avoiding likely false-positive findings – A practical guide. *Biological Reviews, 92*(4), 1941-1968. doi: 10.1111/brv.12315

Frane, A. V. (2015a). Are per-family Type I error rates relevant in social and behavioral science? *Journal of Modern Applied Statistical Methods, 14*(1), 12-23. doi: 10.22237/jmasm/1430453040

Frane, A. V. (2015b). Planned hypothesis tests are not necessarily exempt from multiplicity adjustment. *Journal of Research Practice, 11*(1), article P2. Retrieved from http://jrp.icaap.org/index.php/jrp/article/view/514/417

Frane, A. V. (2016). False discovery rate control is not always a replacement for Bonferroni adjustment (Letter commenting on: J Clin Epidemiol.

2014;67:850-7). *Journal of Clinical Epidemiology, 69*(1), 263. doi: 10.1016/j.jclinepi.2015.03.025

Glickman, M. E., Rao, S. R., & Schultz, M. R. (2014). False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *Journal of Clinical Epidemiology, 67*(8), 850-857. doi: 10.1016/j.jclinepi.2014.03.012

Goeman, J. J., & Solari, A. (2014). Multiple hypothesis testing in genomics. *Statistics in Medicine, 33*(11), 1946-1978. doi: 10.1002/sim.6082

González-Blanch, C., Gleeson, J. F., Cotton, S. M., Crisp, K., McGorry, P. D., & Alvarez-Jimenez, M. (2015). Longitudinal relationships between expressed emotion and cannabis misuse in young people with first-episode psychosis. *European Psychiatry, 30*(1), 20-25. doi: 10.1016/j.eurpsy.2014.07.002

Ha, R. R., & Ha, J. C. (2012). *Integrative statistics for the social and behavioral sciences*. Thousand Oaks, CA: Sage.

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York, NY: John Wiley & Sons.

Huisingh, C., & McGwin, G., Jr. (2012). An analysis of the use of multiple comparison corrections in ophthalmology research. *Investigative Ophthalmology & Visual Science, 53*(8), 4777. doi: 10.1167/iovs.12-10336

Jenkins, T. M., Toosy, A. T., Ciccarelli, O., Miszkiel, K. A., Wheeler-Kingshott, C. A., Henderson, A. P., Kallis, C., Mancini, L., Plant, G. T., Miller, D. H., & Thompson, A. J. (2009). Neuroplasticity predicts outcome of optic neuritis independent of tissue damage. *Annals of Neurology, 67*(1), 99-113. doi: 10.1002/ana.21823

Kim, W. B., Alavi, A., Walsh, S., Kim, S., & Pope, E. (2015). Epidermolysis bullosa pruriginosa: A systematic review exploring genotype-phenotype correlation. *American Journal of Clinical Dermatology, 16*(2), 81-87. doi: 10.1007/s40257-015-0119-7

Marion-Veyron, R., Lambert, M., Cotton, S. M., Schimmelmann, B. G., Gravier, B., McGorry, P. D., & Conus, P. (2015). History of offending behavior in first episode psychosis patients: A marker of specific clinical needs and a call for early detection strategies among young offenders. *Schizophrenia Research, 161*(2-3), 163-168. doi: 10.1016/j.schres.2014.09.078

McKillup, S. (2012). *Statistics explained: An introductory guide for life scientists*. Cambridge, UK: Cambridge University Press. doi: 10.1017/cbo9781139047500

14

Meijer, R. J., & Goeman, J. J. (2016). Multiple testing of gene sets from gene ontology: Possibilities and pitfalls. *Briefings in Bioinformatics, 17*(5), 808-818. doi: 10.1093/bib/bbv091

Miller, R. G. (1981). *Simultaneous statistical inference* (2nd edition). New York, NY: Springer-Verlag. doi: 10.1007/978-1-4613-8122-8

Moran, M. D. (2003). Arguments for rejecting the sequential Bonferroni in ecological studies. *Oikos, 100*(2), 403-405. doi: 10.1034/j.1600-0706.2003.12010.x

Mossaheb, N., Schäfer, M. R., Schlögelhofer, M., Klier, C. M., Cotton, S. M., McGorry, P. D., & Amminger, G. P. (2013). Effect of omega-3 fatty acids for indicated prevention of young patients at risk for psychosis: When do they begin to be effective? *Schizophrenia Research, 148*(1-3), 163-167. doi: 10.1016/j.schres.2013.05.027

Nakagawa, S. (2004). A farewell to Bonferroni: The problems of low statistical power and publication bias. *Behavioral Ecology, 15*(6), 1044-1045. doi: 10.1093/beheco/arh107

O'Connor, M., Harris, J. M., McIntosh, A. M., Owens, D. G. C., Lawrie, S. M., & Johnstone, E. C. (2009). Specific cognitive deficits in a group at genetic high risk of schizophrenia. *Psychological Medicine, 39*(10), 1649-1655. doi: 10.1017/S0033291709005303

Ostendorf, D. M., Lyden, K., Pan, Z., Wyatt, H. R., Hill, J. O., Melanson, E. L., & Catenacci, V. A. (2017). Objectively measured physical activity and sedentary behavior in successful weight loss maintainers. *Obesity, 26*(1), 53-60. doi: 10.1002/oby.22052

Ozcan, T., Bacak, S. J., Zozzaro-Smith, P., Li, D., Sagcan, S., Seligman, N., & Glantz, C. J. (2017). Assessing weight gain by the 2009 Institute of Medicine guidelines and perinatal outcomes in twin pregnancy. *Maternal and Child Health Journal, 21*(3), 509-515. doi: 10.1007/s10995-016-2134-6

Pagano, R. R. (2013). *Understanding statistics in the behavioral sciences* (10th edition). Boston, MA: Cengage.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *BMJ, 316*, 1236-1238. doi: 10.1136/bmj.316.7139.1236

Phillips, A., Fletcher, C., Atkinson, G., Channon, E., Douiri, A., Jaki, T., Maca, J., Morgan, D., Roger, J. H., & Terrill, P. (2013). Multiplicity: Discussion points from the Statisticians in the Pharmaceutical Industry multiplicity expert group. *Pharmaceutical Statistics, 12*(5), 255-259. doi: 10.1002/pst.1584

Racette, L., Boden, C., Kleinhandler, S. L., Girkin, C. A., Liebmann, J. M., Zangwill, L. M., Medeiros, F. A., Bowd, C., Weinreb, R. M., Wilson, M. R., & Sample, P. A. (2005). Differences in visual function and optic nerve structure between healthy eyes of Blacks and Whites. *JAMA Ophthalmology, 123*(11), 1547-1553. doi: 10.1001/archopht.123.11.1547

Rajapakse, T., Griffiths, K. M., Christensen, H., & Cotton, S. (2014). A comparison of non-fatal self-poisoning among males and females, in Sri Lanka. *BMC Psychiatry, 14*(221). doi: 10.1186/s12888-014-0221-z

Roberts, J., Williams, K., Carter, M., Evans, D., Parmenter, T., Silove, N., Clark, T., & Warren, A. (2011). A randomised controlled trial of two early intervention programs for young children with autism: Centre-based with parent program and home-based. *Research in Autism Spectrum Disorders, 5*(4), 1553-1566. doi: 10.1016/j.rasd.2011.03.001

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology, 1*(1), 43-46. doi: 10.1097/00001648-199001000-00010

Rothman, K. J. (2014). Six persistent research misconceptions. *Journal of General Internal Medicine, 29*(7), 1060-1064. doi: 10.1007/s11606-013-2755-z

Rubin, M. (2017). Do *p* values lose their meaning in exploratory analyses? It depends how you define the familywise error rate. *Review of General Psychology, 21*(3), 269-275. doi: 10.1037/gpr0000123

Rutherford, A. (2011). *ANOVA and ANCOVA: A GLM approach* (2nd edition). Hoboken, NJ: John Wiley & Sons. doi: 10.1002/9781118491683

Saville, D. J. (1990). Multiple comparison procedures: The practical solution. *The American Statistician, 44*(2),174-180. doi: 10.2307/2684163

Savitz, D. A. (2003). *Interpreting epidemiologic evidence: Strategies for study design and analysis*. Oxford, UK: Oxford University Press.

Savitz, D. A., & Olshan, A. F. (1995). Multiple comparisons and related issues in the interpretation of epidemiologic data. *American Journal of Epidemiology, 142*(9), 904-908. doi: 10.1093/oxfordjournals.aje.a117737

Scheff, S. W. (2016). *Fundamental statistical principles for the neurobiologist: A survival guide*. London: Academic Press. doi: 10.1016/C2015-0-02471-6

Shulz, K. F., & Grimes, D. A. (2005). Multiplicity in randomised trials I: Endpoints and treatments. *The Lancet, 365*(9470), 1591-1595. doi: 10.1016/S0140-6736(05)66461-6

16

Shulz, K. F., & Grimes, D. A. (2006). *The Lancet handbook of essential concepts in clinical research*. Edinburgh, UK: Elsevier.

Sinclair, J. K., Taylor, P. J., & Hobbs, S. J. (2013). Alpha level adjustments for multiple dependent variable analyses and their applicability – A review. *International Journal of Sports Science and Engineering, 7*(1), 17-20. Retrieved from

http://www.worldacademicunion.com/journal/SSCI/SSCIvol07no01paper03.pdf

Smyth, B. P., James, P., Cullen, W., & Darker, C. (2015). "So prohibition can work?" Changes in the use of novel psychoactive substances among adolescents attending a drug and alcohol treatment service following a legislative ban. *The International Journal of Drug Policy, 26*(9), 887-889. doi: 10.1016/j.drugpo.2015.05.021

Tukey, J. W. (1953). The problem of multiple comparisons. In H. I. Braun (Ed.), *The collected works of John W. Tukey: Multiple comparisons: 1948-1983* (Vol. VIII, pp. 1-300). New York, NY: Chapman & Hall.

van Gils, E. J. M., Veenhoven, R. H., & Hak, E. (2009). Effect of reduced-dose schedules with 7-valent pneumococcal conjugate vaccine on nasopharyngeal pneumococcal carriage in children: A randomized controlled trial. *JAMA, 302*(2), 159-167. doi: 10.1001/jama.2009.975

Westfall, P. (1985). Simultaneous small-sample multivariate Bernoulli confidence intervals. *Biometrics, 41*(4), 1001-1013. doi: 10.2307/2530971

Young, S. S. (2009). *Everything is dangerous: A controversy*. Paper presented at the RAND Statistics Seminar, Pittsburgh, PA. Retrieved from https://www.niss.org/sites/default/files/Young_Safety_June_2008.pdf

Zintzaras, E., & Lau, J. (2008). Synthesis of genetic association studies for pertinent gene-disease associations requires appropriate methodological and statistical approaches. *Journal of Clinical Epidemiology, 61*(7), 634-645. doi: 10.1016/j.jclinepi.2007.12.011