Compromising Outcomes

Peter B. Imrey

Department of Quantitative Health Sciences, Lerner Research Institute and Mellen Center for Multiple Sclerosis Treatment and Research, Neurological Institute, Cleveland Clinic, Cleveland, Ohio; and Cleveland Clinic Lerner College of Medicine of Case Western Reserve University, Cleveland, Ohio

JASN 30: 1147–1150, 2019. doi: https://doi.org/10.1681/ASN.2019010057

Controlled clinical trials are generally regarded as providing the best evidence of therapeutic benefits and harms, because they incorporate powerful tools to control research validity threats: concurrent controls, randomization, masking/blinding of treatments, and sample size predetermination for tolerable false positive and false negative error rates. However, the quality of clinical trial evidence depends on how well trials are executed and reported. The Consolidated Standards for Reporting Trials (CONSORT) guidelines[1] are endorsed by *JASN* as criteria to which authors and editors should aspire to conform. However, a recent review by Chatzimanouil *et al.*[2] raises concerns about adherence to specific CONSORT guidelines in 1996–2016 clinical trial reports in *JASN*, *American Journal of Kidney Diseases* (*AJKD*), *Kidney International* (*KI*), *New England Journal of Medicine*, and *The Lancet*.

Perhaps the most striking of these concerns is lack of clear outcome reporting. Because diseases and treatments have many manifestations, benefit and harm are multidimensional; thus, virtually all trials report multiple outcomes, many repeatedly during follow-up. Despite their ubiquity, multiple outcomes challenge both trial planners and consumers of trial results, because the chances of false positive and false negative scientific conclusions in otherwise well conducted trials often depend crucially on choices among alternative statistical approaches to multiple outcome analyses. The most prevalent approach involves declaration at a trial's outset of a hierarchy of primary, secondary, and tertiary outcomes, with the former acting, in principle, as gatekeeper of an actionable statistical significance claim. However, *ad hoc* reporting and/or opaque reporting about multiple outcomes are increasingly recognized as generating controversy, confusion, and irreproducibility of clinical research findings.[3–24] Unfortunately, despite modest progress from 1996 to 2016, Chatzimanouil *et al.*[2] classify fewer than one half of *JASN* articles that they reviewed from this period as having sufficiently clarified which prespecified outcomes were initially defined as primary and secondary end points, including how and when assessed (figure 3B, outcome a in the work by Chatzimanouil *et al.*[2]). Moreover, almost no reports among those reviewed, and none whatsoever from *JASN*, *AJKD*, and *KI*, were classified as having clearly identified, explained, or disclaimed changes in prespecified outcomes during the trial (figure 3B, outcome b in the work by Chatzimanouil *et al.*[2]).

The finding of Chatzimanouil *et al.*[2] is disturbing in the context of current reproducibility concerns due to the critical role played by outcome predetermination/prioritization in managing and describing a trial's false positive and false negative error rates. This paper provides a brief conceptual overview of the technical "multiplicity" problem, accepted solutions, common evasions, and the need to address such evasions by increasing awareness and expectations of transparency.

As illustration, consider two parallel-arm clinical superiority trials comparing drugs X and Y. Suppose that Trial A records occurrences of a single efficacy outcome (*e.g.*, rate of eGFR decline) annually during 4 follow-up years, and Trial B records cumulative occurrences of each of four dichotomous efficacy outcomes (eGFR decline >10 ml/min per 1.73 m^2, AKI, dialysis onset, and all-cause mortality) once at the end of follow-up. In Trial A, if the four annual treatment differences are tested separately with, for simplicity, one-sided 2.5%-level hypothesis tests consistent with the Food and Drug Administration (FDA) standards for drug superiority trials or in Trial B, if each outcome is similarly compared separately between treatments, then the chance of falsely concluding that drug X is superior to drug Y if the drugs are truly equivalent is negligible—below one in 16,000 ($4\times0.975\times0.025^3+0.025^4=0.0061\%$)—if, to claim overall superiority, significance is required for at least 3 years (Trial A) or three outcomes (Trial B) and yearly results (Trial A) or outcomes (Trial B) are assumed to be statistically independent (a simplifying, if unrealistic, assumption for illustration). However, if only one significant time or outcome is required, then under the same assumption, this chance is 9.63% ($1-(0.975)^4$), virtually quadrupling the nominal 2.5%

Published online ahead of print. Publication date available at www.jasn.org.

Correspondence: Dr. Peter B. Imrey, Department of Quantitative Health Sciences JJN3, Cleveland Clinic, 9500 Euclid Avenue, Cleveland, OH 44195. Email: imreyp@ccf.org

false positive rate. This problem worsens commensurately with more observation times and/or outcomes. Dependencies among times and outcomes alter the specific numbers but not the problem.

Although worst in "omics" studies involving tens of thousands to millions of simultaneous significance tests, this "multiple comparison problem" has been known to statisticians for a century and under discussion in the context of clinical trials for at least four decades.[25] An extensive statistical literature provides several types of solutions within the clinical research literature's usual "frequentist" hypothesis testing statistical framework, each with variations adapted to differing clinical and statistical situations. These are summarized broadly in Figure 1 and in more detail by statistical texts,[26,27] a recent European Medicines Agency guidance,[28] a US FDA draft guidance,[29] and a recent journal special section.[30–34] Broadly speaking, one may (*1*) restructure hypotheses to use fewer tests as with an overall analysis of variance (ANOVA) test or Pearson chi-squared test comparing *k* proportions or (*2*) redefine and control error rates in groups of tests rather than at the individual test level (*e.g.*, as the probability of any false positives among a collection of tests ["familywise error rate"] or the fraction of false positives among all positive results ["false discovery rate"]). Analyses from a Bayesian statistical framework offer additional attractive solutions but may introduce other interpretive difficulties. Regardless, the available methods for handling multiple outcomes are powerful and flexible, and when used well, they can limit hypothesis testing errors and consequently, also limit incorrect scientific findings from purely random variation.

They cannot, however, increase information in the study sample. The price of successful random error control, as for methods of statistical error control in general, is to pick a method and execute it as planned. In principle, this largely requires adherence to the initially planned hierarchy of outcomes and specific statistical procedure(s), because it is mathematical analyses and/or computational

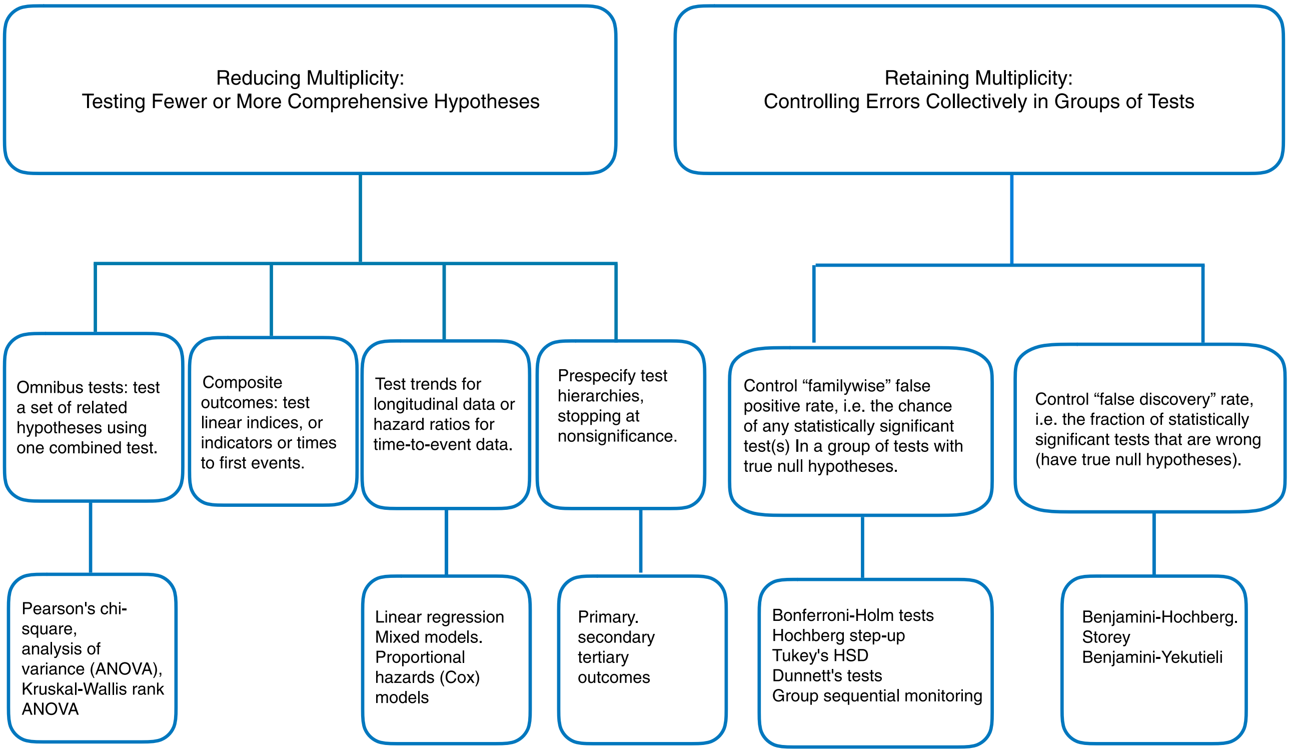
Reducing Multiplicity:
Testing Fewer or More Comprehensive Hypotheses
Retaining Multiplicity:
Controlling Errors Collectively in Groups of Tests
Omnibus tests: test a set of related hypotheses using one combined test.
Composite outcomes: test linear indices, or indicators or times to first events.
Test trends for longitudinal data or hazard ratios for time-to-event data.
Prespecify test hierarchies, stopping at nonsignificance.
Control "familywise" false positive rate, i.e. the chance of any statistically significant test(s) In a group of tests with true null hypotheses.
Control "false discovery" rate, i.e. the fraction of statistically significant tests that are wrong (have true null hypotheses).
Pearson's chi-square, analysis of variance (ANOVA), Kruskal-Wallis rank ANOVA
Linear regression Mixed models. Proportional hazards (Cox) models
Primary. secondary tertiary outcomes
Bonferroni-Holm tests Hochberg step-up Tukey's HSD Dunnett's tests Group sequential monitoring
Benjamini-Hochberg. Storey Benjamini-Yekutieli

Figure 1. At least six strategies are available for error control of multiple simultaneous hypothesis tests within the clinical research literature's prevalent "frequentist" approach to statistical inference. The group on the left indicates four ways by which multiple comparison concerns in analyses of clinical trial data may be assuaged by reducing the number of formal statistical hypothesis tests that are simultaneously considered: omnibus testing of composite hypotheses, formation and testing of composite outcomes, comparisons of trends or rates of change rather than individual times points in analyses of longitudinal repeated measures outcomes, and formation of outcome hierarchies in which decisions predicated on statistical significance of lower outcomes in the hierarchy require statistical significance of outcomes higher in the hierarchy. The group on the right indicates two ways by which multiple comparison concerns can be managed, without reducing the number of tests, by controlling error rates defined in terms of results of groups of tests rather than individual tests: control of the "familywise error rate" (*i.e.*, of the chance that any member of the group will yield a statistically significant false positive result when all null hypotheses tested are true) and control of the "false discovery rate" (*i.e.*, of the fraction of positive test results resulting from a larger group of tests that are false positives). All of these methods are frequently used, and combinations of them may be required to adequately control false positive inferences from clinical trials. HSD, honestly significant difference.

simulations based upon these that establish the study's claimed error control properties. These error properties may be more or less sensitive to changes in a study's analytic plan, but no established rules of thumb predict their sensitivity to omissions or substitutions among outcomes or rearrangement of an outcome hierarchy. *Post hoc* changes in handling multiple outcomes, including drawing conclusions about treatment efficacy from secondary or tertiary outcomes of a trial with nonsignificant primary outcome, abandoning or reshuffling components of an outcome hierarchy, or declaiming that significance tests are exploratory while drawing conclusions as if they were confirmatory, thus preserve the form but vitiate the technical meaning of statistical significance, with its intrinsic linkage to a specified false positive error rate.

This is a quandary for investigators, because the technical requirement of fidelity to a previously specified approach, while controlling undesirable "fishing expeditions," also restricts health scientists' natural proclivities to apply aspects of what is learned from data to their interpretation, and restricts statistical scientists' natural desires to ensure that technical assumptions of analytic modeling processes are realistic and that analysis efficiently uses the information in the data. The quandary understandably causes discomfort, increased by the perception that journal acceptance rates are tightly linked to statistically significant *P* values. However, wisely accepting this discomfort honors the Nobel physicist Richard Feynman's "first principle" of scientific integrity: "that you must not fool yourself—and you are the easiest person to fool. So you have to be very careful about that."[35] As humans, we are supremely talented at finding threads—"stories"—to describe our experiences, scientific theories being one type of such stories. However, we are less talented at assessing probability and relative plausibility against a background of chance variation. We tend to be suckers for clues, and there are many possible stories, relatively few of which convey great insight.[36]

Many reviews have documented that outcome omissions, substitutions, reprioritizations, and inventions, often unannounced, appear with disturbing frequency in clinical research reports.[3–24] Even when motivated by the sincere desire to capture all possible insights from hard-earned, expensive data, such changes are inherently undesirable and often harmful. They breach the implicit contract between investigators, funders, oversight bodies, and participating patients by invalidating important aspects of the trial rationale under which the study was approved and funded, and patients were consented and enrolled. Although the comments above have stressed issues with efficacy outcomes and false positive errors, similar considerations apply to safety outcomes and errors in identifying safety signals.

This does not mean that deviations from prespecified outcomes and analyses are always scientifically wrong and impermissible. It does mean, however, that such changes should always require clear rationale and be identified clearly in a paper's main text, with rationale fully visible to reviewers, editors, and readers, preferably with results from the original plan presented for easy comparison. This allows readers to assess the arguments for and against any notable alterations in scientific findings from those on the basis of the initial plan. Transparency is no panacea. Here, however, as in other circumstances where principles may collide, it makes the information needed to understand and rationally debate an issue more available; discourages individuals from arbitrary decisions hard to publicly defend; illuminates disagreements and sometimes, paths to consensus as well; and fosters trust, the lifeblood of science. Such transparency is consistent with the spirit and letter of CONSORT and all other widely accepted research reporting guidelines. However, increased understanding of the dangers of compromising outcomes by investigators; heightened alertness by reviewers, editors, and readers to inconsistencies of reported outcomes and their statistical treatments with study protocols and analysis plans; and strengthened journal policies requiring transparency when addressing such inconsistencies, will all be needed for transparent, CONSORT-compliant outcome reporting to become the norm.

ACKNOWLEDGMENTS

Dr. Imrey is grateful to *JASN* editorial colleagues and submitting authors, whose efforts have stimulated this commentary.

DISCLOSURES

Dr. Imrey reports nonfinancial support from the Task Force on Design and Analysis of Oral Health Research and the Foundation for the Task Force on Design and Analysis of Oral Health Research all outside the submitted work.

FUNDING

This work was partially supported by the National Center for Advancing Translational Sciences, grant U54TR001960; the National Institute of Arthritis and Musculoskeletal and Skin Diseases, grants R01AR068342 and R01AR074131; and the Agency for Health Care Research and Quality, grant R01HS024277.

REFERENCES

1. Schulz KF, Altman DG, Moher D; CONSORT Group: CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *Trials* 11: 32, 2010
2. Chatzimanouil MKT, Wilkens L, Anders HJ: Quantity and reporting quality of kidney research. *J Am Soc Nephrol* 30: 13–22, 2019
3. Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG: Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *JAMA* 291: 2457–2465, 2004
4. Chan AW, Krleza-Jeric K, Schmid I, Altman DG: Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *CMAJ* 171: 735–740, 2004
5. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E, et al.: Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS One* 3: e3081, 2008

6. Mathieu S, Boutron I, Moher D, Altman DG, Ravaud P: Comparison of registered and published primary outcomes in randomized controlled trials. *JAMA* 302: 977–984, 2009
7. Kirkham JJ, Dwan KM, Altman DG, Gamble C, Dodd S, Smyth R, et al.: The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ* 340: c365, 2010
8. Hannink G, Gooszen HG, Rovers MM: Comparison of registered and published primary outcomes in randomized clinical trials of surgical interventions. *Ann Surg* 257: 818–823, 2013
9. Rosenthal R, Dwan K: Comparison of randomized controlled trial registry entries and content of reports in surgery journals. *Ann Surg* 257: 1007–1015, 2013
10. Dwan K, Gamble C, Williamson PR, Kirkham JJ; Reporting Bias Group: Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. *PLoS One* 8: e66844, 2013
11. Dwan K, Altman DG, Clarke M, Gamble C, Higgins JP, Sterne JA, et al.: Evidence for the selective reporting of analyses and discrepancies in clinical trials: A systematic review of cohort studies of clinical trials. *PLoS Med* 11: e1001666, 2014
12. Page MJ, McKenzie JE, Kirkham J, Dwan K, Kramer S, Green S, et al.: Bias due to selective inclusion and reporting of outcomes and analyses in systematic reviews of randomised trials of healthcare interventions. *Cochrane Database Syst Rev* 10: MR000035, 2014
13. Roest AM, de Jonge P, Williams CD, de Vries YA, Schoevers RA, Turner EH: Reporting bias in clinical trials investigating the efficacy of second-generation antidepressants in the treatment of anxiety disorders: A report of 2 meta-analyses. *JAMA Psychiatry* 72: 500–510, 2015
14. Fleming PS, Koletsi D, Dwan K, Pandis N: Outcome discrepancies and selective reporting: Impacting the leading journals? *PLoS One* 10: e0127495, 2015
15. Jones CW, Keil LG, Holland WC, Caughey MC, Platts-Mills TF: Comparison of registered and published outcomes in randomized controlled trials: A systematic review. *BMC Med* 13: 282, 2015
16. Pandis N, Fleming PS, Worthington H, Dwan K, Salanti G: Discrepancies in outcome reporting exist between protocols and published oral health Cochrane systematic reviews. *PLoS One* 10: e0137667, 2015
17. Coronado-Montoya S, Levis AW, Kwakkenbos L, Steele RJ, Turner EH, Thombs BD: Reporting of positive results in randomized controlled trials of mindfulness-based mental health interventions. *PLoS One* 11: e0153220, 2016
18. Ioannidis JPA, Caplan AL, Dal-Ré R: Outcome reporting bias in clinical trials: Why monitoring matters. *BMJ* 356: j408, 2017
19. Lancee M, Lemmens CMC, Kahn RS, Vinkers CH, Luykx JJ: Outcome reporting bias in randomized-controlled trials investigating antipsychotic drugs. *Transl Psychiatry* 7: e1232, 2017
20. Jones PM, Chow JTY, Arango MF, Fridfinnson JA, Gai N, Lam K, et al.: Comparison of registered and reported outcomes in randomized clinical trials published in anesthesiology journals. *Anesth Analg* 125: 1292–1300, 2017
21. Goldacre B, Drysdale H, Powell-Smith A, Dale A, Ioan M, Eirion S, et al.: The COMpare Trials Project, 2016. Available at: www.compare-trials.org. Accessed May 29, 2019
22. Hopewell S, Witt CM, Linde K, Icke K, Adedire O, Kirtley S, et al.: Influence of peer review on the reporting of primary outcome(s) and statistical analyses of randomised trials. *Trials* 19: 30, 2018
23. Lee S, Khan T, Grindlay D, Karantana A: Registration and outcome-reporting bias in randomized controlled trials of distal radial fracture treatment. *JBJS Open Access* 3: e0065, 2018
24. Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, et al.: Reducing waste from incomplete or unusable reports of biomedical research. *Lancet* 383: 267–276, 2014
25. Tukey JW: Some thoughts on clinical trials, especially problems of multiplicity. *Science* 198: 679–684, 1977
26. Dmitrienko A, Tamhane AC, Bretz F: Multiple Testing Problems in Pharmaceutical Statistics, Boca Raton, FL, CRC Press, 2010
27. Westfall PH, Tobias RD, Wolfinger RD: Multiple Comparisons and Multiple Tests Using SAS, 2nd Ed., Cary, NC, SAS Institute Inc., 2011
28. European Medicines Agency: Guideline on Multiplicity Issues in Clinical Trials EMA/CHMP/44762/2017, 2017. Available at: https://www.ema.europa.eu/documents/scientific-guideline/draft-guideline-multiplicity-issues-clinical-trials_en.pdf. Accessed May 29, 2019
29. FDA: FDA Draft Guidance for Industry. Multiplicity Endpoints in Clinical Trials, 2017. Available at: http://www.fda.gov/media/102657/download. Accessed June 7, 2019
30. Dmitrienko A, D'Agostino RB: Editorial: Multiplicity issues in clinical trials. *Statist Med* 36: 4423–4426, 2017
31. Chuang-Stein C, Li J: Changes are still needed on multiple co-primary endpoints. *Statist Med* 36: 4427–4436, 2017
32. Sankoh AJ, Li H, D'Agostino RB: Composite and multicomponent end points in clinical trials. *Statist Med* 36: 4437–4440, 2017
33. Snappin S: Some remaining challenges regarding multiple endpoints in clinical trials. *Statist Med* 36: 4441–4445, 2017
34. Dmitrienko A, Millen B, Lipkovich I: Multiplicity considerations in subgroup analysis. *Statist Med* 36: 4446–4454, 2017
35. Feynman RP: Cargo cult science. In: *Surely You're Joking, Mr. Feynman: Adventures of a Curious Character*, edited by Feynman RP, New York, W.W. Norton, 1984, pp 338–446
36. Kahneman D: Thinking, Fast and Slow, New York, Farrar, Strauss and Giroux, 2011