Epidemiology 4

Multiplicity in randomised trials I: endpoints and treatments

Kenneth F Schulz, David A Grimes

Lancet 2005; 365: 1591–95

Family Health International, PO Box 13950, Research Triangle Park, NC 27709 USA
(K F Schulz PhD, D A Grimes MD)

Correspondence to:
Dr Kenneth F Schulz
KSchulz@fhi.org

Multiplicity problems emerge from investigators looking at many additional endpoints and treatment group comparisons. Thousands of potential comparisons can emanate from one trial. Investigators might only report the significant comparisons, an unscientific practice if unwitting, and fraudulent if intentional. Researchers must report all the endpoints analysed and treatments compared. Some statisticians propose statistical adjustments to account for multiplicity. Simply defined, they test for no effects in all the primary endpoints undertaken versus an effect in one or more of those endpoints. In general, statistical adjustments for multiplicity provide crude answers to an irrelevant question. However, investigators should use adjustments when the clinical decision-making argument rests solely on one or more of the primary endpoints being significant. In these cases, adjustments somewhat rescue scattershot analyses. Readers need to be aware of the potential for under-reporting of analyses.

Many analytical problems in trials stem from issues related to multiplicity. Investigators usually address the issues responsibly; however, others ignore or remain oblivious to their ramifications. Put colloquially, some researchers torture their data until they speak. They examine additional endpoints, manipulate group comparisons, do many subgroup analyses, and undertake repeated interim analyses. Difficulties usually manifest at the analysis phase because investigators add unplanned analyses. Literally thousands of potential comparisons can emanate from one trial, in which case many significant results would be expected by chance alone. Some statisticians propose adjustments in response, but unfortunately those adjustments frequently create more problems than they solve.[1]

Multiplicity problems stem from several sources. Here, we address multiple endpoints and multiple treatments. In the next article[2] we address subgroup and interim analyses. The perspectives on multiplicity are contentious and complex.[3–6] In proposing approaches to handle multiplicity, any position alienates many (panel 1). Multiplicity issues stir hot debates.[10]

The issue

Multiplicity portends troubles for researchers and readers alike for two main reasons. First, investigators should report all analytical comparisons implemented. Unfortunately, they sometimes hide the complete analysis, handicapping the reader's understanding of the results. Second, if researchers properly report all comparisons made, statisticians proffer statistical adjustments to account for multiple comparisons. Investigators need to know whether they should use such adjustments, and readers whether to expect them.

Multiplicity can increase the overall error in significance testing. The type 1 error (α), under the hypothesis of no association between two factors, indicates the probability of the observed association from the data at hand being attributable to chance. It advises the reader of the likelihood of a false-positive conclusion.[11] The problem emerges when multiple

See *Lancet* 2005; **365:** 1348–53

Panel 1: **Divergent views on statistical adjustments for multiplicity**

Some statisticians favour adjustments for multiple comparisons, whereas others disagree.
"Several recent publications show that the multiple comparisons debate is alive and well. I . . . observe that it is hard to see views such as the following being reconciled . . ."[7]
"No adjustments are needed for multiple comparisons."[4]
"Bonferroni adjustments are, at best, unnecessary and, at worst, deleterious to sound statistical inference."[1]
". . . Type I error accumulates with each executed hypothesis test and must be controlled by the investigators."[8]
"Methods to determine and correct type 1 errors should be reported in epidemiologic and public health research investigations that include multiple statistical tests."[9]

independent associations are tested for significance. If d=the number of comparisons, then the probability that at least one association will be found significant is $1–(1–\alpha)^d$. Frequently, investigators in medical research set α at 0·05. Thus, if they test ten independent associations, assuming the universal null hypothesis of no association in all ten, the probability of at least one significant result is 0·40, ie, $(1–[1–0·05]^{10})$. Stated alternatively, the cumulative chance of at least one false-positive result out of the ten comparisons is 40%. Nevertheless, the probability of a false positive for every single comparison remains 0·05 (5%) whether one or a million are tested.[4]

A proposed statistical solution

Most statisticians would recommend reducing the number of comparisons as a solution to multiplicity. Given many tests, however, some statisticians recommend making adjustments such that the overall probability of a false-positive finding equals α after making d comparisons in the trial. Authors usually attribute the method to Bonferroni and simply state that to test comparisons in a trial at α, all comparisons should be performed at the α/d significance level, not at the α level.[5,12] Thus, for an α of 0·05, with ten comparisons, every test would have to be significant at the 0·005 level. Analogously, some investigators retain the same individual α threshold but multiply every observed p value by d.[10,13] Thus, with ten comparisons, an observed p=0·02 from a trial would yield an adjusted p=0·20. Of note, the Bonferroni adjustment inflates β error thereby reducing statistical power.[1]

Bonferroni adjustment, however, usually addresses the wrong hypothesis.[1,6] It assumes the universal null hypothesis which, simply defined, tests that two groups are identical for all the primary endpoints investigated versus the alternative hypothesis of an effect in one or more of those endpoints. That usually poses an irrelevant question in medical research. Clinically, a similar idea 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."[1] Indeed, Rothman wrote: "To entertain the universal null hypothesis is, in effect, to suspend belief in the real world, and thereby to question the premises of empiricism."[4]

Drug regulation with the need for clear dichotomous answers appropriately drives much of the activity in multiplicity adjustments. Adjustments fit the hypothesis-testing paradigm—approval or no approval—needed for drug regulation. In most published medical research, however, we encourage the presentation of interval estimation (eg, relative risks with confidence intervals) for effects rather than just hypothesis testing (just a p value).[14] Moreover, we suggest that the decision-making intent in most medical research discourages multiplicity adjustments.

Multiple endpoints

Although the ideal approach for the design and analysis of randomised controlled trials relies on one primary endpoint, investigators typically examine more than one. The most egregious abuse with multiplicity arises in the data-dredging that happens behind the scenes and remains unreported. Investigators analyse many endpoints, but only report the favourable significant comparisons. Failure to note all the comparisons actually made is unscientific if unwitting and fraudulent if intentional. "*Post hoc* selection of the end-point with the most significant treatment difference is a deceitful trick which invariably overemphasizes a treatment difference."[13] Investigators must halt this deceptive practice.

Researchers should restrict the number of primary endpoints tested. They should specify a priori the primary endpoint or endpoints in their protocol. Focusing their trial increases the simplicity of implementation and the credibility of results. Furthermore, they should follow their protocol for their analysis. Deviations for data-dredging can be condoned, but should be clearly labelled as explorations and fully reported. Disappointingly, trial reports frequently contain examinations of endpoints not included in the trial protocol but ignore planned primary analyses from the protocol.[15] Safeguards to ensure that investigators have followed the protocol (such as *The Lancet*'s protocol acceptance track and asking for protocols for all randomised controlled trials) provide assistance, but more extensive registering and publishing of protocols makes sense. Lastly, investigators must report all the comparisons made.[16,17]

Statistical adjustments for multiple endpoints might sabotage interpretation. For example, suppose investigators undertook a randomised controlled trial of a new antibiotic compared with a standard antibiotic for prevention of febrile morbidity after hysterectomy. They designated fever the primary outcome, and the results showed a 50% reduction (relative risk 0·50 [95% CI 0·25–0·99]; p=0·048). Note the significant result. Alternatively, suppose they had designated two primary endpoints: wound infection and fever. As typically happens in trials, the endpoints are highly correlated. So in addition to the 50% reduction in fever, the trial also found a 52% decrease in wound infection (0·48 [0·24–0·97]; p=0·041). From some statisticians' viewpoints, investigators should correct for multiple comparisons by, for example, multiplying every p value by the number of comparisons made—ie, $0·048\times2=0·096$ and $0·041\times2=0·082$. Both p values adjust to $>0·05$; thus the trial would be indeterminate (negative).

Seasoned clinical trialists, however, look at these results quite differently. The wound infection result enhances rather than debases the first result on fever. Clinicians understand biologically that the two endpoints are highly related. Adding the second endpoint on wound infection and observing similar results lends credence to the observed reduction in febrile morbidity. That adjustments

would abolish the basic finding defies logic.[1] Doing so would somewhat resemble a doctor finding an abnormally low amount of haemoglobin in a patient but no longer judging it worthy of treatment because they also obtained an abnormal packed-cell volume (haematocrit).

Indeed, some statisticians would agree with not using formal adjustments for multiplicity in the aforementioned example. Even those predisposed to such adjustments recommend against them under certain delineated clinical decision-making scenarios.[3] If an investigator proposes to claim treatment effect if all the endpoints are significant or if most (defined in the protocol) are significant, then they assert that no adjustment for multiple endpoints is necessary.[3]

Furthermore, the Bonferroni adjustment, advocated most frequently for multiplicity, is an overcorrection at best. Moreover, it can be a severe overcorrection when the endpoints are associated with one another,[3,13] which is generally the case. Overcorrecting for p values hampers interpretation of results. The adjustment for multiple comparisons "mechanizes and thereby trivializes the interpretive problem, and it negates the value of much of the information in large bodies of data".[4] Clinical insights remain important. Investigators need to focus on the smallest number of endpoints that makes clinical sense and then report results on all endpoints tested. If more than one primary endpoint exists, they should discuss whether additional endpoints reinforce or detract from the core findings. Formal adjustments for multiplicity frequently obscure rather than enhance interpretation.

Composite endpoints

Composite endpoints alleviate multiplicity concerns.[18] A composite endpoint happens if any one of the prospectively defined components of the composite takes place. For example, a composite cardiovascular endpoint would happen if myocardial infarction, stroke, or cardiovascular death arose. If designated a priori as the primary outcome, the composite obviates the multiple comparisons associated with testing of the separate components. Moreover, composite outcomes usually lead to high event rates thereby increasing power or reducing sample size requirements. Not surprisingly, investigators frequently use composite endpoints.[18]

However, interpretational difficulties sometimes arise. For example, aspirin produced an 18% reduction (relative risk 0·82 [95% CI 0·70–0·96]) in the above-defined composite endpoint of cardiovascular events (myocardial infarction, stroke, or cardiovascular death), a seemingly worthwhile result.[19] However, a secondary look at the separate components revealed a 44% decrease in myocardial infarction, a 22% increase in stroke, and virtually no effect on cardiovascular death. That 18% reduction seems meaningless in view of the lack of beneficial effect on the relatively more important outcomes of death and stroke.[19] Composite endpoints frequently lack clinical relevancy.[20] Thus, composite endpoints address multiplicity and generally yield statistical efficiency at the risk of creating interpretational difficulties.

***Panel 2:* A role for multiarm trials in medical research**

Multiarm trials are fairly common in the medical literature. A search of parallel designed randomised controlled trials indexed on PubMed in 2000 revealed 25% with more than two arms. Of those, 62% had three arms, 26% had four, and 12% had more than four (Altman DG, personal communication).

The preponderance of material in clinical trial textbooks addresses two-arm trials. Furthermore, eminent researchers have strongly recommended against more than two-arms: "A positive result is more likely, and a null result is more informative, if the main comparison is of only two treatments, these being as different as possible."[22] The argument against multiarm trials mainly centres on trial power. Published trials typically have inadequate power.[23] Given a finite number of potential participants, the argument holds that adding arms only further dilutes power. Although we sympathise with this argument, multiarm trials might not only be attractive in some circumstances but also be more efficient.

For example, imagine an instance whereby a standard treatment exists and two new potentially effective therapies have materialised. A two-arm approach dictates a comparison of a new with standard and then probably an additional trial of the other new with a group from the first trial. In general, the overall study size and cost would be greater with this sequential two-arm approach than with one multiarm trial. Multiarm trials sometimes make sense. Furthermore, multiarm trials do not necessarily raise methodological concerns. They can eliminate selection bias just like two-arm trials. Although they tend to be more complex to undertake and analyse, that complexity frequently yields commensurate gains in information.

Multiple treatments (multiarm trials)

Addressing multiplicity from multiple treatments is a more tractable problem than from multiple endpoints. First, investigators can avert multiple tests by one global test of significance across comparison groups[13]—eg, comparing A *vs* B *vs* C in a three-arm trial—or by modelling a dose-response relation.[21] Second, and perhaps most importantly, researchers have less opportunity to data-dredge on many treatments and not report them. While they easily can add more endpoints for analysis, they would have difficulty adding treatments in a trial. They theoretically could implement a multigroup trial and then only report the favourable group comparisons, but little evidence exists for that practice. We suspect that readers of a trial report usually see all the treatments implemented. Indeed, multiarm trials have an important role in medical research (panel 2).

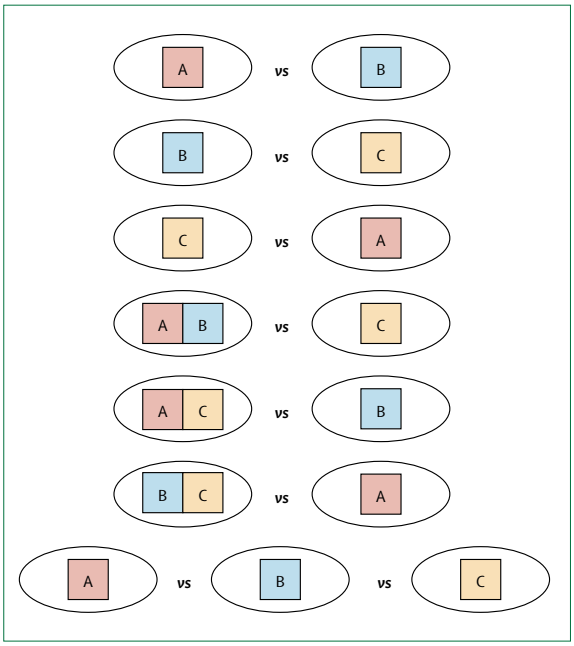

Figure: **At least seven possible comparisons from a three-arm trial**

However, the situation is not entirely sanguine. What readers of a journal article might not see are all the different comparisons among the treatment groups. For example, with a three-arm trial, at least seven possible analyses emerge (figure). With more than three arms the potential comparisons explode. Obviously, investigators should specify a priori the comparisons intended.

With multiarm trials, as mentioned earlier, a frequently recommended approach entails undertaking one global test across all treatments. However, some methodologists believe such tests are of limited use because they do not identify which treatments are different and because of limited power to detect genuine differences.[13] Many multiarm trials are designed for direct comparison with controls.[13] Thus, investigators should plan the comparisons intended, limit the number, and document them in the protocol.

Adjustments for multiple comparisons generally need not have a role in multigroup trials. Similar to the above argument for multiple endpoints, clinicians usually find the addition of a group to a trial enhances rather than diminishes informativeness. For example, in the randomised controlled trial described earlier, comparing a new antibiotic with standard treatment for prevention of fever after hysterectomy, investigators might add a treatment group with a 300 mg dose to a trial of a 200 mg dose of the same antibiotic. The results showed a 40% reduction for the 200 mg dose (relative risk 0·60 [95% CI 0·37–0·98]; p=0·044). Note the significant result. The 300 mg dose expectedly yields a similar result, a 45% decrease in fever (0·55 [0·31–0·98]; p=0·041). The simple adjustment approach for multiple comparisons involves multiplying every p value by the number of comparisons made—ie, 0·044×2=0·088 and 0·041×2=0·082. With adjustment, the effects become non-significant at the 0·05 level and thus indeterminate (negative).

Again, however, trialists interpret these results quite differently. The result for 300 mg augments rather than degrades the result for 200 mg on fever. Clinicians expect similar results biologically. They would seriously distrust adjustment that abolishes those significant results. Adjusting p values, particularly with related treatment groups, does not aid in interpreting the results of the trial.

With multiple treatments, investigators sometimes use a prioritised sequence of tests.[24] For example, investigators might decide on the 300 mg new antibiotic versus standard treatment as the priority test and, if that comparison is significant, only then proceed to the 200 mg comparison. Such procedures address multiplicity without adjustments.[24] Again, formal adjustments for multiplicity usually complicate rather than enlighten.

The role of adjustments for multiplicity

Sometimes formal adjustments for multiplicity are inescapable. An obvious example would arise with certain decision-making criteria in submissions to a regulatory agency for drug approval. If the sponsor specifies more than one primary endpoint and proposes to claim treatment effect if one or more are significant, investigators should adjust for multiplicity.[3] Furthermore, the same principle extends to all investigators whose decision-making intent is to claim an effect based on any one of a number of endpoints being positive.

Adjustments might also be indicated in a multiarm trial in which investigators plan a scattershot analysis. For example, in a four-arm trial (treatments A, B, C, and D), they intend on claiming an effect for A if any one of the following comparisons yielded significant results: A versus B, A versus C, A versus D, A versus B+C, A versus B+D, A versus C+D, or A versus B+C+D. The best recourse might be a multiplicity adjustment.

In general, when prudence indicates multiplicity adjustments, trials tend to be poorly and diffusely designed. An adjustment for multiplicity merely partly salvages credibility. Moreover, even when adjustment becomes appropriate, implementation becomes difficult. Bonferroni adjustments are generally recommended, usually because of their simplicity. However, other adjustment strategies sometimes perform better.[3,25] Depending on the correlation among the endpoints, simulation experiments display wide variability in α error and power of various multiplicity adjustment strategies.[3] These comparative assessments help, but still clear-cut choices prove elusive. The adjustments usually provide crude answers.

What readers should look for

Readers should expect the researchers to report all the endpoints analysed and treatments compared. Assessing whether they reported them all is usually difficult.

Access to the protocol would be helpful but is usually impossible. We urge greater access to protocols. Poor, incomplete reporting, however, frequently renders readers helpless to know the complete analysis undertaken by the investigators. Reporting according to the CONSORT statement obviates these difficulties.[16,17]

Readers should expect the primary endpoint or endpoints to be specified, with other analyses being labeled as exploratory. In lieu of direct statements, search for indirect indications. If the primary endpoint remains unclear, hopefully the authors provided a statistical power analysis that indicates the primary endpoint.

Readers should expect some interpretation if authors make multiple comparisons. If authors went overboard and reported results on 15 endpoints with one being significant they should display appropriate caution. If multiple comparisons yield multiple effects, authors should address the internal consistency of the results. Most importantly, transparent reporting of all comparisons allows readers to come to their own interpretations.

If a trial report specifies a composite endpoint, the components of the composite should be in the well-known pathophysiology of the disease. The researchers should interpret the composite endpoint in aggregate rather than as showing efficacy of the individual components. However, the components should be specified as secondary outcomes and reported beside the results of the primary analysis.[18]

In general, readers need not expect corrections for multiplicity. For most trials, adjustments for multiplicity lack substance and prove unhelpful. An exception might include a medical research article with an argument that rests solely on one or more of the primary endpoints being significant, essentially the test of the universal null hypothesis. An adjustment for multiplicity somewhat rescues such scattershot analyses.

Conflict of interest statement

We declare that we have no conflict of interest.

Acknowledgments

We thank David L Sackett, Douglas G Altman, and Willard Cates for their helpful comments on an earlier version of this report.

References

1 Perneger TV. What's wrong with Bonferroni adjustments. *BMJ* 1998; **316:** 1236–38.
2 Schulz KF, Grimes DA. Multiplicity in randomised trials II: subgroup and interim analyses. *Lancet* (in press).
3 Sankoh AJ, D'Agostino RB Sr, Huque MF. Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues. *Stat Med* 2003; **22:** 3133–50.
4 Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology* 1990; **1:** 43–46.
5 Westfall P, Bretz F. Multiplicity in clinical trials: encyclopedia of biopharmaceutical statisitics, 2nd edn. New York: Marcel Dekker, 2003: 666–73.
6 Savitz DA, Olshan AF. Multiple comparisons and related issues in the interpretation of epidemiologic data. *Am J Epidemiol* 1995; **142:** 904–08.
7 Altman DG. Statistics in medical journals: some recent trends. *Stat Med* 2000; **19:** 3275–89.
8 Moye LA. P-value interpretation and alpha allocation in clinical trials. *Ann Epidemiol* 1998; **8:** 351–57.
9 Ottenbacher KJ. Quantitative evaluation of multiplicity in epidemiology and public health research. *Am J Epidemiol* 1998; **147:** 615–19.
10 Altman DG. Practical statistics for medical research. London: Chapman and Hall, 1991.
11 Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. *Lancet* 2005; **365:** 1348–53.
12 Friedman L, Furberg C, DeMets D. Fundamentals of clinical trials. St Louis: Mosby, 1996.
13 Pocock SJ. Clinical trials: a practical approach. Chichester: Wiley, 1983.
14 Sterne JA, Davey Smith G. Sifting the evidence: what's wrong with significance tests? *BMJ* 2001; **322:** 226–31.
15 Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 2004; **291:** 2457–65.
16 Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports or parallel-group trials. *Lancet* 2001; **357:** 1191–94.
17 Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001; **134:** 663–94.
18 Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials: greater precision but with greater uncertainty? *JAMA* 2003; **289:** 2554–59.
19 Pocock SJ. Clinical trials with multiple outcomes: a statistical perspective on their design, analysis, and interpretation. *Control Clin Trials* 1997; **18:** 530–45.
20 Meinert CL. Clinical trials: design, conduct, and analysis. New York: Oxford University Press, 1986.
21 Senn S. Statistical issues in drug development. Chichester: John Wiley and Sons, 1997.
22 Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient I: introduction and design. *Br J Cancer* 1976; **34:** 585–612.
23 Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* 1994; **272:** 122–24.
24 Bauer P, Chi G, Geller N, et al. Industry, government, and academic panel discussion on multiple comparisons in a "real" phase three clinical trial. *J Biopharm Stat* 2003; **13:** 691–701.
25 Hsu JC. Multiple comparisons: theory and methods. New York: Chapman and Hall, 1996.