



Non-adjustment for multiple testing in multi-arm trials of distinct treatments: Rationale and justification

Richard A Parker and Christopher J Weir

Abstract

There is currently a lack of consensus and uncertainty about whether one should adjust for multiple testing in multi-arm trials of distinct treatments. A detailed rationale is presented to justify non-adjustment in this situation. We argue that non-adjustment should be the default starting position in simple multi-arm trials of distinct treatments.

Keywords

Multiple testing, multi-arm clinical trial, family-wise error rate, type-I error, multiplicity, alpha adjustment

Introduction

The multi-arm trial is an efficient trial design which has been applied in many different settings.^{1–4} Recent examples of this trial design include the MS-SMART trial⁴ in which three treatments were compared against a common control group in multiple sclerosis patients, and the TAME trial comparing four novel treatments for severe acute malnutrition in children.⁵ All multi-arm trials by their nature will involve multiple testing due to the multiple treatment comparisons, leading to an increased probability of at least one false-positive error across all hypotheses of interest. Formally, this is called the family-wise type-I error rate (FWER), which is defined as the probability of making at least one false-positive conclusion among all the multiple hypotheses being tested.⁶ Multi-arm trial designs that utilise a common control group are also associated with an increased risk of multiple false-positive errors; that is, the probability of K or more errors occurring among the hypotheses of interest (where $K > 1$).⁶ Most multiplicity adjustment methods control the FWER. It is also possible to additionally control the multiple error rate⁶ or the expected number of false-positive conclusions.⁷ In particular, the well-known Bonferroni correction not only controls the FWER regardless of how many null hypotheses are true, but also controls the expected number of false-positive conclusions.⁷

If multi-arm trials have treatments that are strongly related (e.g. treatments with different doses), then there is widespread agreement that a multiplicity adjustment is necessary.^{1–3,8,9} However, for multi-arm trials of

distinct treatments, there is a lack of consensus and uncertainty about whether one should adjust for multiple testing in this setting.^{1,3,6,8} This lack of consensus may imply that the arguments against multiplicity adjustment are weak or insignificant, but we would argue that this is false. In this article, we aim to provide a strong rational basis to support non-adjustment in multi-arm trials of distinct treatments.

The statistical basis of multiple testing adjustment

The concept of a global null hypothesis

If we are using a multiple testing correction that (as a minimum) aims to reduce or control the FWER, then this implies that we wish to control the false-rejection rate of a global null hypothesis. The global null hypothesis consists of the intersection hypothesis of all null hypotheses that we are interested in (called a ‘family’ of null hypotheses). For $i = 1, \dots, m$ null hypotheses, we can write the following

$$H_0 : H_{0,1} \cap H_{0,2} \cap \dots \cap H_{0,m}$$

Edinburgh Clinical Trials Unit, Usher Institute, The University of Edinburgh, Edinburgh, UK

Corresponding author:

Richard A Parker, Edinburgh Clinical Trials Unit, Usher Institute, The University of Edinburgh, Level 2, Nine Edinburgh BioQuarter, 9 Little France Road, Edinburgh EH16 4UX, UK.
Email: Richard.Parker@ed.ac.uk

In the case of multi-arm trials, each of the null hypotheses $H_{0,1}, H_{0,2}, \dots, H_{0,m}$ could indicate that there is no difference between treatment and control, for example. The alternative hypothesis would then be the global union hypothesis

$$H_A : H_{A,1} \cup H_{A,2} \cup \dots \cup H_{A,m}$$

This is called ‘union-intersection testing’¹⁰ and implies that if any of the individual null hypotheses are rejected at an unadjusted alpha significance level, then so would the global null hypothesis.

The global intersection null hypothesis is more likely to be rejected as the number of individual null hypotheses increases with multiple testing, and so multiple testing adjustment is designed to control this inflated probability of rejection. If treatments are distinct and we are interested in individual treatment versus control comparisons, however, then it is difficult to see how the concept of formulating a global intersection null hypothesis could be relevant. If the global intersection null hypothesis is not relevant, then neither is the FWER.

Nevertheless, a valid counter-argument is that although we might not be *explicitly* interested in the global intersection null hypothesis, the final analysis results may be reported and interpreted in ways that *implicitly* correspond to an underlying global null hypothesis. This is easily seen if we consider the example of a four-arm trial, with three of the arms consisting of the same treatment given at different doses, and the fourth arm placebo. In this case, if only one of the treatment doses shows a significant effect relative to placebo, then we would conclude that the treatment is effective. It is therefore recommended to control the FWER in this situation so that the probability of incorrectly concluding that the treatment is effective is below a pre-specified level. In this case, the overall treatment efficacy conclusion is the same regardless of which of the individual doses showed treatment efficacy, and therefore, we are implicitly assuming that the rejection of a global intersection hypothesis is clinical meaningful.

In multi-arm trials of distinct treatments, however, we are usually interested in each of the treatment comparisons individually (e.g. whether drug A is effective compared to placebo) and their associated false-positive error rate, often referred to as the *comparison-wise error rate*. The comparison-wise error rate is not inflated by multiple testing.¹⁰ Indeed, the expected proportion of incorrectly rejected hypotheses will not exceed the significance level used in the individual tests.¹⁰

The single claim of effectiveness

Some authors have emphasised the importance of making a ‘single claim of effectiveness’ or defining a ‘clinical win criteria’ as determining the need for multiple testing

adjustment.^{6,7,9,10} Howard et al.⁶ for example, have suggested that an FWER adjustment is only necessary if ‘assessing multiple hypotheses within a multi-arm trial has increased the chance of making a single claim of effectiveness’. This decision criteria requires careful definition and consideration because a statement such as ‘treatment A is effective’ (assuming no other treatments are effective) could be regarded as a ‘single claim of effectiveness’, even though it has resulted from individual interpretation of comparison-wise hypothesis tests. As above, comparison-wise testing does not require a multiplicity adjustment. Clearly, the term ‘single claim of effectiveness’ must refer to making a conclusion such as ‘the treatment is effective’ as in the dose–response example, which will occur whenever any of the multiple hypotheses is rejected. If we name specific treatments (e.g. state that ‘treatment A is effective’) in our clinical efficacy conclusion, which we invariably would do when interpreting results from multi-arm trials of distinct treatments, we have already severed the connection between the global intersection hypothesis at the study level and our clinical efficacy conclusion. Therefore, researchers should be clear that it is the *content* of the clinical efficacy statement that matters and how it is determined, not how many statements of effectiveness are stated or how many conclusions given in a final manuscript. Accordingly, the definition of ‘single claim of effectiveness’ and its specific rationale should be included in any overall justification of a multiple testing adjustment.

Rationale for non-adjustment

Issues with multiplicity adjustment

The idea of controlling the FWER to minimise the risk of a study recommending an ineffective treatment for use in clinical practice sounds appealing from the statistical perspective. Nevertheless, it does not accord with the fact that within medicine or public health, multi-arm trials are usually focussed on the effects of individual treatments. Knowledge of the effectiveness of at least one of the treatments is not by itself clinically useful information – the clinician needs to know which treatments are effective so that they can make an impact on clinical practice or lead to further investigation. If one of the treatments was successful, then we need to know its specific name and pharmacological properties. How many other treatments were tested alongside it in the same trial is irrelevant information if those treatments were distinct. It therefore seems paradoxical to control overall errors resulting from entire projects or studies when our main focus in multi-arm trials is usually in the individual treatment comparisons themselves. After all, as others have commented, if the treatments were evaluated in separate trials, there

would be no expectation that the researcher would need to make a multiplicity adjustment.^{6,11}

More generally, it is not always clear which hypotheses should be included in the ‘family’ when calculating the FWER in multiplicity adjustment. The word ‘family’ in FWER does not have a universally consistent definition and its definition needs to be carefully considered for each multiple testing application.^{6,12} In our experience ‘family’ tends to be defined at the level of the study protocol or the collection of primary hypotheses that are assessed in a final analysis. However, it is difficult to rationalise why it should be defined at the study level (e.g. instead of across all trials which include the same treatment versus control comparisons), other than relying on convention and pragmatism. Howard et al. recommend that ‘family’ is taken to encompass all hypotheses which contribute to a ‘single statement of effectiveness’;⁶ but as we have seen, the very definition of ‘single statement of effectiveness’ is itself uncertain and open to differing interpretations.

Furthermore, setting the type I error rate of the overall study to be below a certain level obscures the importance of setting individually appropriate error rate thresholds according to context and expected costs, which is strongly recommended in the literature.^{12–14} If we are prioritising errors at the study level over errors in individual treatment comparisons then we are saying in effect that the impact of the study is more important than the impact of the clinical treatments, which is contrary to common sense.

In addition, treatment recommendations arising from a study are often based on both clinical and statistical significance¹⁵ and may be based on secondary outcomes and/or safety outcomes as well as primary outcomes. Therefore, the interpretative environment used in clinical trials is often much more complex than that in which a multiple testing adjustment implies. Statistical evidence from a set of primary null hypotheses may not be sufficient in itself to lead to clinical impact or substantial cost from a type I error.¹⁵ There is a need to consider the totality of the evidence in favour of a treatment.¹⁶

Another argument against multiplicity adjustment is that it usually over-complicates interpretation of the study results.¹⁷ Unadjusted and adjusted results may lead to different conclusions, and adjusted results are not easy to interpret individually because the degree of adjustment will depend on the underlying global hypothesis and how many other treatment comparisons have been made.

There is also the danger that by focussing too much on controlling type I error, we overlook the type II error rate (failing to reject the null hypothesis even though it is false). The consequence of reducing the type I error rate is that the type II error rate is

increased.^{2,13,18,19} This is why performing multiplicity adjustment is unsupported in exploratory or early phase studies where type II error is important.

The confirmatory trials argument

It has been suggested that for confirmatory multi-arm trials, the FWER should be controlled but not for exploratory trials.^{1,11} There is some logic in making the distinction between exploratory and confirmatory. For confirmatory trials, or trials in which we wish to make definitive conclusions to settle controversy, the cost of a false-positive finding is much greater because the treatment is likely to gain wide acceptance or proceed to be used in clinical practice. However, just because a multi-arm trial is labelled as ‘confirmatory’ does not mean that any of the issues associated with multiple testing adjustment have been resolved. In particular, adjusting for multiplicity is still inconsistent with how multi-arm trials of distinct treatments are usually interpreted.

Some might argue that it still makes sense to minimise the danger in interpreting a result as definitive and ‘confirmatory’ if only one treatment arm is significant among many tested. But we argue that this problem should not be addressed by multiple testing – it should be addressed instead by replication and multiple experiments to confirm the results, or by reducing the individual α significance levels themselves. Indeed, regulatory agencies usually require ‘statistically compelling and clinically relevant’ results for the licencing of medicines, which typically involves replication of the results in at least one other trial.²⁰ If there was only one pivotal trial conducted, then the results have to be particularly compelling, with a strong pharmacological rationale and strong statistical significance.²⁰ Anecdotally this usually involves setting a statistical significance level of 0.25%, equating to two independent trials being significant at 5%. In any case, if a type I error rate is of great concern for a given treatment, then this should be addressed by the individual α -levels themselves – not indirectly via controlling the overall FWER.

Consideration of some further objections

Some researchers might argue that it is better to guard against certain types of risky interpretation in an overall study rather than relying on the fact that individual treatments will be interpreted in isolation. However, it is not easy to predict how results will be interpreted or influence subsequent behaviour,¹⁵ and inappropriate adjustment has ethical consequences because it leads to a larger sample size, lower statistical power or a combination of these.⁶ The Bonferroni adjustment in particular imposes an extreme penalty in this regard.

Another potential objection is that if the multi-arm trial has a common control group, then we have to adjust for multiple testing anyway because the treatment comparisons are related in this way. However, Howard et al.⁶ have demonstrated that this concept is false. The FWER is not increased in multi-arm trials with a common control group compared to what it would be if the treatments were tested in independent trials. Therefore, multiplicity adjustment is not required purely on the basis of sharing control data.⁶

Some might argue that the more treatment arms in a multi-arm trial, the more likely we are to require adjustment.¹¹ Again, as for confirmatory studies, this does not solve the real issues covered above. Indeed, the counter-arguments to adjustment may strengthen when the number of arms assessed increases, because the multiplicity adjustment becomes more stringent and the type II error is further inflated. This could be compensated by increasing the sample size, but this is not without cost since it may increase the overall participant burden, study duration, or financial cost of the study.

Discussion

In general, non-adjustment for multiple testing is a valid and statistically defensible approach for multi-arm trials of distinct treatments. We do accept however, that in some circumstances, for example, when the same treatments are evaluated in multiple subgroups,⁹ adjustment for multiple testing is appropriate. We also recognise that for confirmatory trials of investigational medicinal products (IMPs), multiplicity adjustment may be required by some regulators, for example, the European Medicines Agency.^{3,21} We do not attempt to argue that adjustment is never required. Our main purpose in writing this article is to provide a robust case as to why the decision to not to proceed to perform adjustment is valid as a general proposition. Any decision to adjust or not adjust should be thought through carefully with proper consideration of the study objectives, study design, and analysis.^{3,22}

In this article, we have only considered multiplicity adjustment of the simple case of multi-arm trials with distinct treatments. If multi-arm trials have arms of related treatments, have multiple primary outcomes, multiple subgroups within treatment arms,⁹ or include interim analyses, then it is likely that some form of multiplicity adjustment will be required. In particular, if an overall positive treatment efficacy conclusion does not depend on *which* of the individual hypothesis tests are statistically significant, and does not require all the individual tests to be significant, then a multiplicity adjustment is recommended.

An argument often presented in favour of non-adjustment is that if independent trials were done, then no multiple testing correction would be performed in this case.^{2,3,6,8} In fact, the reasons for non-adjustment go deeper than this. Although adjusting for multiple testing enables control of the FWER at the study level and may make sense theoretically, we would argue that it does not make good sense from the perspective of clinical practice and can lead to difficulties with interpretation. It also tends to be logically incompatible with the main clinical questions of interest.¹⁵

Most of the arguments against multiple testing adjustment presented in this article are not new. As far back as 1990, Rothman questioned the validity of posing a global null hypothesis if there is no strong justification for doing so.^{11,18} In his 1998 paper, Perneger¹⁹ argued against the use of the Bonferroni correction in general to correct for multiple testing. He questioned why a type I error rate should change depending on the number of tests performed.¹⁹ In 1996, Cook and Farewell¹⁵ wrote that 'In particular, a concern is that testing strategies are frequently adopted with the aim of controlling the experimental type I error rate without considering how this relates to the questions of main interest'. We think this concern is still pertinent today. Automatic and unthinking adjustment for multiple testing without thoughtful consideration is dangerous and potentially unethical. The decision about whether to adjust for multiple testing is not an abstract exercise, but one that may have major ethical implications.⁶ For example, if an unnecessary multiple testing adjustment is planned and incorporated into the sample size calculation, then this may require the study to have a much larger sample size than needed, which is a waste of time and resources. This is why we disagree with sweeping statements promoting the more widespread use of multiplicity adjustment.²³ Careful thought is needed as to whether multiplicity adjustment is necessary in each specific circumstance.



Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors were partly supported in this work by NHS Lothian via the Edinburgh Clinical Trials Unit.

ORCID iDs

Richard A Parker  <https://orcid.org/0000-0002-2658-5022>
Christopher J Weir  <https://orcid.org/0000-0002-6494-4903>

References

1. Wason JM, Stecher L and Mander AP. Correcting for multiple-testing in multi-arm trials: is it necessary and is it done? *Trials* 2014; 15(1): 364.
2. Freidlin B, Korn EL, Gray R, et al. Multi-arm clinical trials of new agents: some design considerations. *Clin Cancer Res* 2008; 14(14): 4368–4371.
3. Juszczak E, Altman DG, Hopewell S, et al. Reporting of multi-arm parallel-group randomized trials: extension of the CONSORT 2010 statement. *JAMA* 2019; 321(16): 1610–1620.
4. Chataway J, De Angelis F, Connick P, et al. Efficacy of three neuroprotective drugs in secondary progressive multiple sclerosis (MS-SMART): a phase 2b, multiarm, double-blind, randomised placebo-controlled trial. *Lancet Neurol* 2020; 19(3): 214–225.
5. Kelly P, Bell L, Amadi B, et al. TAME trial: a multi-arm phase II randomised trial of four novel interventions for malnutrition enteropathy in Zambia and Zimbabwe – a study protocol. *BMJ Open* 2019; 9: e027548.
6. Howard DR, Brown JM, Todd S, et al. Recommendations on multiple testing adjustment in multi-arm trials with a shared control group. *Stat Methods Med Res* 2018; 27(5): 1513–1530.
7. Jaki T and Parry A. Why are two mistakes not worse than one? A proposal for controlling the expected number of false claims. *Pharm Stat* 2016; 15(4): 362–367.
8. Li G, Taljaard M, Van den Heuvel ER, et al. An introduction to multiplicity issues in clinical trials: the what, why, when and how. *Int J Epidemiol* 2017; 46(2): 746–755.
9. Stallard N, Todd S, Parashar D, et al. On the need to adjust for multiplicity in confirmatory clinical trials with master protocols. *Ann Oncol* 2019; 30(4): 506–509.
10. Dmitrienko A, Bretz F, Westfall PH, et al. Multiple testing methodology. In: Dmitrienko A, Tamhane AC and Bretz F (eds) *Multiple testing problems in pharmaceutical statistics*. 1st ed. Chapman & Hall/CRC Biostatistics Series, Boca Raton, Florida, U.S.A. 2010. Pages 35–41.
11. Proschan MA and Waclawiw MA. Practical guidelines for multiplicity adjustment in clinical trials. *Control Clin Trials* 2000; 21(6): 527–539.
12. Berger VW. On the generation and ownership of alpha in medical studies. *Control Clin Trials* 2004; 25(6): 613–619.
13. Grieve AP. How to test hypotheses if you must. *Pharm Stat* 2015; 14(2): 139–150.
14. Parker RA. Overcoming obstacles to deriving sample size calculations: experiences of a biostatistician. *SAGE Res Meth Cases* 2020. DOI: 10.4135/9781529731699.
15. Cook RJ and Farewell VT. Multiplicity considerations in the design and analysis of clinical trials. *J R Stat Soc Series A* 1996; 159: 93–110.
16. Ioannidis JP. Why most published research findings are false. *PLOS Med* 2005; 2(8): e124.
17. Schulz KF and Grimes DA. Multiplicity in randomised trials I: endpoints and treatments. *Lancet* 2005; 365(9470): 1591–1595.
18. Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology* 1990; 1(1): 43–46.
19. Perneger TV. What's wrong with Bonferroni's adjustment? *BMJ* 1998; 316: 1236–1238.
20. European Medicines Agency. Committee for proprietary medical products (CPMP): points to consider on application with 1. meta-analyses; 2. one pivotal study, https://www.ema.europa.eu/en/documents/scientific-guideline/points-consider-application-1meta-analyses-2one-pivotal-study_en.pdf (2001, accessed 20 March 2020).
21. European Medicines Agency. Guideline on multiplicity issues in clinical trials [draft], <https://www.ema.europa.eu/en/multiplicity-issues-clinical-trials> (2017, accessed 20 March 2020).
22. European Medicines Agency. *ICH E9 statistical principles for clinical trials*, <https://www.ema.europa.eu/en/ich-e9-statistical-principles-clinical-trials#current-version-section> (1998, accessed 20 March 2020).
23. Cristea IA and Ioannidis JPA. P values in display items are ubiquitous and almost invariably significant: a survey of top science journals. *PLoS ONE* 2018; 13(5): e0197440.