International Journal of Epidemiology, 2017, 746–755
doi: 10.1093/ije/dyw320
Advance Access Publication Date: 26 December 2016
Education Corner

Education Corner

An introduction to multiplicity issues in clinical trials: the what, why, when and how

Guowei Li,[1,2,*] Monica Taljaard,[3,4] Edwin R Van den Heuvel,[5,6] Mitchell AH Levine,[1,2,7] Deborah J Cook,[1,2,7] George A Wells,[3,8] Philip J Devereaux[1,7,9] and Lehana Thabane[1,2,9]

[1]Department of Clinical Epidemiology and Biostatistics, [2]St Joseph's Healthcare Hamilton, McMaster University, Hamilton, ON, Canada, [3]Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, ON, Canada, [4]Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, ON, Canada, [5]Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands, [6]Department of Epidemiology, University Medical Center Groningen, Eindhoven, The Netherlands, [7]Department of Medicine, McMaster University, Hamilton, ON, Canada, [8]Department of Medicine, University of Ottawa, Ottawa, ON, Canada and [9]Population Health Research Institute, Hamilton Health Sciences, McMaster University, Hamilton, ON, Canada

*Corresponding author. Clinical Epidemiology and Biostatistics, McMaster University, St Joseph's Healthcare Hamilton, 501–25 Charlton Avenue East, Hamilton, ON L8N 1Y2, Canada. E-mail: lig28@mcmaster.ca

Accepted 10 October 2016

Abstract

In clinical trials it is not uncommon to face a multiple testing problem which can have an impact on both type I and type II error rates, leading to inappropriate interpretation of trial results. Multiplicity issues may need to be considered at the design, analysis and interpretation stages of a trial. The proportion of trial reports not adequately correcting for multiple testing remains substantial. The purpose of this article is to provide an introduction to multiple testing issues in clinical trials, and to reduce confusion around the need for multiplicity adjustments. We use a tutorial, question-and-answer approach to address the key issues of why, when and how to consider multiplicity adjustments in trials. We summarize the relevant circumstances under which multiplicity adjustments ought to be considered, as well as options for carrying out multiplicity adjustments in terms of trial design factors including Population, Intervention/Comparison, Outcome, Time frame and Analysis (PICOTA). Results are presented in an easy-to-use table and flow diagrams. Confusion about multiplicity issues can be reduced or avoided by considering the potential impact of multiplicity on type I and II errors and, if necessary pre-specifying statistical approaches to either avoid or adjust for multiplicity in the trial protocol or analysis plan.

Key words: Multiplicity adjustment, trial, experiment-wise error rate, type I error

Key Messages

- The proportion of clinical trial reports not adequately correcting for multiple testing remains substantial.
- Multiplicity-related issue considerations may be needed in terms of trials design factors including Population, Intervention/Comparison, Outcome, Time frame and Analysis (PICOTA) in clinical trials.
- We use a tutorial, question-and-answer approach to address the key issues of why, when and how to consider multiplicity adjustments in clinical trials.

Introduction

Multiplicity issues are not uncommon in randomized controlled trials. Multiplicity refers to the potential inflation of the type I error rate as a result of multiple testing, for example due to multiple subgroup comparisons, comparisons across multiple treatment arms, analysis of multiple outcomes, and multiple analyses of the same outcome at different times. A type I error refers to erroneously rejecting the null hypothesis, where the probability of a type I error is commonly referred to as the significance level of the trial. A major implication of multiple testing is that the overall significance level of the trial may need to be adjusted to account for multiplicity.[1] The proportion of trial reports not adequately correcting for multiple testing in the literature remains substantial.[2–4] For example, one study evaluated multi-arm trials published in four major medical journals (the *British Medical Journal*, *Lancet*, *New England Journal of Medicine* and *PLoS Medicine*) between January 2012 and December 2012. It found that among 39 multi-arm confirmatory trials, only 46% performed multiplicity adjustments.[4] This is despite many educational articles[1,5,6] and textbooks[7,8] addressing this topic in practice. The purpose of this article is to provide an introduction to multiple testing adjustments in clinical trials, and to reduce the confusion around the need to adjust for multiplicity. Before any treatment code is broken in a clinical trial, of primary importance is a detailed predefined statistical analysis plan. We use a tutorial-style question-and-answer approach to address the key issues of why, when and how to consider adjustments for multiple testing in trials. Results are presented in easy-to-use tables and flow diagrams to mitigate the burden of reading and understanding, especially for novice researchers.

Why do we need to consider multiple testing adjustments?

When a set of hypotheses are tested simultaneously within the same study, the overall type I error rate (i.e. the probability of rejecting at least one null hypothesis given that all nulls are in fact true) is increased, potentially resulting in an increased risk of a false-positive finding. For instance, if we have five independent or related true null hypotheses, each tested simultaneously at a nominal significance level of $\alpha = 5\%$ (where α refers to the probability of a type I error), the true type I error over all the tests is 23%. This can be easily calculated using the binomial probability distribution, i.e. $P(\text{at least one siginficant result}) = 1 - (1 - \alpha)^k$, where k is the number of tests. Therefore in the case of five tests, if we do not control for multiple testing we will have a 23% chance to obtain at least one significant result when indeed the null hypothesis is true, whereas authors and readers may believe that the type I error rate is maintained at the level of 5%. The probability of making a false-positive finding in multiple testing (in this case, 23%) is also called the experiment-wise or family-wise error rate. If adequate adjustments are not made in multiple testing, findings may be misleading. Besides the increased risk of spurious statistical significance, multiplicity also has important implications for sample size determination and interpretation of study results.[1,9] Therefore we need to consider multiplicity adjustments in designing, analysing and interpreting trials.

Frequently asked questions (FAQs): when should we consider multiplicity adjustments?

What if we have more than two study arms?

One common multi-arm trial design compares multiple experimental interventions with one control arm. This multi-arm design can improve efficiency by reducing the sample size over that required for separate trials, or increase statistical power for the same sample size.[2,4,10] One systematic review reported that among all the randomized controlled trials published in 2009, the proportion of multi-arm designs was 17.6%, reflecting the increased popularity of multi-arm trials.[11] However, when multiple comparisons are made in multi-arm trials, multiplicity adjustments may need to be considered to avoid an increase in the type I error rate. Next, we describe some of the considerations involved in the case of multiple study arms.

Is the trial of exploratory or confirmatory nature?

Exploratory trials often occur earlier in the development of a new intervention (e.g. phase I or phase II trials and some

pilot trials). The key difference between an exploratory and confirmatory trial is that the latter is designed to seek a definitive answer to a specified hypothesis with the findings intended to be used for final decision making, including the licensing of treatments.[12] Whereas findings from an exploratory trial will have to be tested in further trials, results from a confirmatory trial can address a pre-specified key hypothesis for generating evidence to inform decision making.[5] Note that the results of trials designed as confirmatory, depending upon their findings, may also require confirmation or refutation in future trials. An individual trial may also have both confirmatory and exploratory aspects. For example, most confirmatory trials may include further exploratory analyses which can be used to explain or support the trial findings and for suggesting further hypotheses for later research. Another difference between confirmatory and exploratory trials is that confirmatory trials are usually designed to answer the research questions with specified sample size determined in the study protocol, whereas exploratory trials may not be enforced to meet a specified sample size requirement. In a confirmatory clinical trial, it is generally required to carefully consider multiplicity adjustment in a predefined statistical analysis plan, report all the relevant findings transparently and provide appropriate interpretations.[4,13] In contrast, multiplicity adjustments in exploratory trials may not be required, because any positive findings from exploratory trials should undergo additional testing before changing clinical practice.[4,5] Multiplicity adjustments may not be necessary in exploratory trials, but acknowledging the implications of multiplicity is also important to help interpret the trial results.

Are the treatments (arms) distinct or related?

There is a consensus in the literature that multiplicity adjustments are required if the different treatment arms are related.[4] For instance, if a trial evaluates different dosages or regimens of a treatment compared with the same control arm, then adequate multiple testing adjustments should be performed. The underlying rationale is that if any of the null hypotheses being tested is rejected, it would potentially lead to a recommendation in favour of the new treatment. For example, a phase III randomized controlled trial was conducted to explore the effect of addition of docetaxel to two platinum regimens (i.e. docetaxel plus cisplatin, or docetaxel plus carboplatin) on survival compared with a same control arm of standard first-line chemotherapy (i.e. vinorelbine and cisplatin) in chemotherapy-naive patients with advanced non small-cell lung cancer.[14] Because both the two treatment arms used the same regimen (docetaxel) and added a platinum agent as an adjunct (cisplatin or carboplatin), they were related with each other. Therefore, a multiplicity adjustment was required in this trial.[15] However, there is currently no consensus on the need for multiplicity adjustment in a multi-arm trial with distinct treatment arms, according to findings from a recent review.[4] Multiplicity adjustments may be of lesser importance in the case of distinct treatment arms. For instance, it may be less important to adjust for multiplicity in a smoking cessation study which compares two different intervention arms (educational training, medical intervention) with no intervention. The rationale is that this situation is analogous to running two separate trials under the same protocol; in the case of a separate trial, it would not be required to adjust for the other trial being conducted simultaneously.

Are findings summarized in one conclusion?

It is generally recognized that multiplicity should be controlled if the findings are summarized in one single conclusion for a multi-arm trial.[1,5] The reason is that such finding would be based on comparisons that are implicitly correlated due to the shared control arm, and thus summarizing the findings in one single conclusion essentially involves testing multiple connected primary hypotheses simultaneously. Therefore, all the comparisons included in the single conclusion are regarded as one experiment or family of connected comparisons.[15] For example, if the global objective is to assess whether two new treatments (T1, T2) are 'both superior to the control arm (C)', it is necessary to adjust for multiplicity - because the global comparison essentially involves testing two connected primary hypotheses regarding the superiority of each of the two treatments against the control.[5]

What if there are multiple outcomes?

Clinical trials often assess multiple outcomes (or 'endpoints') such as symptoms, blood test results, side effects, quality of life, or death, to try to maximize the usefulness of information from a costly trial. For example in a cardiovascular trial, outcomes of interest may include hospitalization, stroke, heart failure, myocardial infarction, cardiac arrest, disability and death. If we test each of the individual outcomes separately at a nominal 5% level and obtain any significant difference, the probability of a spurious claim of significance is higher than the anticipated 5%.[2,16,17]

To avoid inflation of the type I error rate, several solutions have been proposed.[2,5,16–18] The first option is to identify one single outcome as the primary outcome and to treat the remaining outcomes as secondary in the study design. There is no need to adjust for multiplicity when there is a single primary outcome, as findings for secondary

outcomes are considered subsidiary and exploratory, rather than confirmatory.

A second potential solution that has been proposed is to use a composite outcome by including all the outcomes, for example based on the time-to-first-event principle. Composite outcomes have several advantages, including that they allow one to choose a combination of relevant outcomes and avoid issues related to competing risks, and they can improve statistical power over using any single outcome, among others.[2] However, concerns have been raised about the use of composite outcomes, including difficulty with respect to interpreting a significant difference in a composite outcome comprising many outcomes, and the underlying implication that all the individual outcomes involved in a composite are of similar importance to patients.[19–21]

A third solution that has been suggested is to conduct a global measure test for the multiple outcomes by adding up the standardized effect sizes for all the individual outcomes with weights reflecting different importance of outcomes, and then testing the summed effect size.[17] A disadvantage of this approach is that determining appropriate weights for effect sizes of the individual outcomes can be challenging because of the need to account for correlations between the individual outcomes.[1,17] Another similar solution termed 'win ratio' has been proposed recently, in which the approach could prioritize the more important component of the composite (e.g. cause-specific mortality) and use the number of pairs in which the patient on new treatment 'won' divided by the numbers of pairs 'lost' to produce a win ratio.[22] The win ratio approach has gained increasing popularity in cardiovascular research; however, concerns have been raised including: failure to employ actual survival times from randomization to event occurrence; the potential impairment of the randomization process due to the matching procedure; and the potential exclusion of large percentages of the less important components preceding the more important components in the win ratio calculation settings, among others.[22,23]

What if we conduct multiple interim analyses?

Multiple analyses for the same outcomes at different fixed time points are often required as interim analyses for accumulating data to monitor trials over the long term.[5,24] They aim to determine whether to stop the trial early if the new treatments are significantly superior to the control arm or cause more adverse events. Conducting multiple analyses entails repeated use of the same data, thereby increasing the type I error rate. It is therefore necessary to consider multiplicity adjustments to account for interim analyses of the same outcome at different time points. The extensive literature for interim analyses focuses on methods for group-sequential designs.[25–27] Note that it is usually required to choose a stringent stopping rule for pre-specified interim monitoring, to obtain a significance level of close to 0.05 in the final analysis.[24,28–31] For instance, it may be adequate to select a very small *P*-value of < 0.001 as stopping rule in interim analyses.[30–32] The HOPE (Heart Outcomes Prevention Evaluation) trial aimed to assess the effect of ramipril (an angiotensin-converting enzyme inhibitor) compared with placebo on the composite of myocardial infarction, stroke or death from cardiovascular causes in patients at high risk of cardiovascular events.[33] For the stopping rules in interim analyses, the trial used a *P*-value of ≤ 0.00003 during the first half and a *P*-value of ≤ 0.002 in the second half of the trial. By doing so, the trial retained the *P*-value of close to 0.05 (i.e.,$0.05\text{-}0.00003\text{-}0.002 = 0.048$) for the final analysis.[33] Alternatively, identification of stopping rules can be based on the crude estimate of treatment effect (e.g. O'Brien and Fleming's method[34] and Kittelson and Emerson's method[35]), the normalized Z statistic and the fixed sample *P*-value (e.g. Wang and Tsiatis's method[36] and Pocock's method[25]), or the error spending function (e.g. Lan and DeMets's approach[37] and Kim and DeMets's method[38]), which is summarized in a tutorial in detail by Emerson *et al.*[39]

What if we have repeated measurements for the same outcomes?

It is not uncommon that the same outcome is assessed repeatedly over time on the same participants. Repeated testing of the same outcome at different times will generally lead to an inflated type I error rate; thus it is required to consider multiplicity adjustments for individual measurement for the same outcome.[1,40,41] Clinical trials using repeated measures have to be longitudinally analysed, in order to take trends over time into account. The primary interest in such trials is usually the over-time difference between the study arms. For example, researchers may repeatedly measure outcomes such as blood pressure, drug clearance fraction, depression or pain scores, whether admitted to hospitals in the intervention and control groups over time. Multiplicity adjustments may have to be considered for between-subject effects (e.g. differences between treatment groups), within-subject effects (e.g. within-subject differences over time) or both (e.g. difference between treatment groups and within-subject differences over time).[5] Alternatively, to avoid multiplicity adjustment, one potential solution is to analyse all the repeated measurements in a single model after the data collection is complete, using either repeated measures analysis

of variance (ANOVA) or a mixed-effect model in which the group by time interaction is assessed for statistical significance.[5] However, this solution may not allow evaluation of the treatment effect comparisons at each time point (contrasts). Testing for such contrasts can be achieved in longitudinal analyses by having a model with a treatment effect parameter at each follow-up time and averaging over the treatment effect parameters, or by incorporating a time-treatment interaction in a model. Other options may include using the measurements at the last time point (or other predefined fixed time point) as the primary outcome and treating measurements at other time points as secondary outcomes. By doing so, it may be less necessary to adjust for multiplicity because there is one single primary outcome. Likewise, another potential solution for avoiding multiplicity adjustment may be to create a single summary score of the repeated measurements, such as area under the curve.

What if we have multiple secondary outcomes?

As mentioned above in the first solution for multiple outcomes in, findings for secondary outcomes are usually treated as exploratory results. Any definitive finding for secondary outcomes may require further confirmatory studies to support them. Therefore it is less necessary to adjust for multiple testing for multiple secondary outcomes.

What if we run multiple subgroup analyses for the same outcomes?

It is not uncommon to undertake subgroup analyses in clinical trials to determine whether the overall trial finding applies to all eligible patients or whether there is any difference in effect of interventions between subgroups defined on, for example, sex, age, presence of comorbidity or severity of illness.[42] Authors presenting subgroup analyses need to carefully consider whether any multiplicity issues would arise. Multiplicity adjustments are required if predefined subgroup analyses are specified for the following reasons: (i) to confirm biological plausibility (because differences in baseline characteristics may have a sound biological basis to support subgroup claims); (ii) to confirm reasonable existing hypotheses (in which the hypotheses are formulated based on previous good quality evidence); and (iii) to show subgroup effects for supporting decision making in target populations (by seeking a definitive answer to specified subgroup hypotheses to inform decision making). Subgroup analyses satisfying the three aforementioned conditions are considered as having a confirmatory nature, thereby requiring multiplicity adjustments.[5,24,43,44] Whereas many trials pre-specify subgroup analyses which are considered supportive to examine treatment consistency, others conduct *a posteriori* subgroup analyses after the finalized protocol and/or data collection have been completed. *Post hoc* subgroup analyses are usually considered to be of an exploratory nature.[5] Thus, *post hoc* subgroup analyses after the study is done can only be used to generate hypotheses, rather than making strong inferences.[24] Consequently there is less need to adjust for multiplicity in *a posteriori* subgroup analyses for the same outcomes. However, of note, the overall trial findings are always better estimates of treatment effect than *a posteriori* subgroup results, regardless of whether the trial is considered exploratory or confirmatory. Reasons include lack of power to detect subgroup effects, the caveats of data dredging which may lead to false subgroup effects, the risk of spurious significant findings purely because of chance, and scepticism due to lack of consistency in the case of qualitative interactions (i.e. one treatment is beneficial in some subgroups but harmful in others).[42,45,46]

What if we conduct multiple sensitivity analyses for the same outcomes?

Sensitivity analyses in trials are usually performed to evaluate whether the main results are consistent under a range of different assumptions, thereby assessing the robustness of the key findings. Sensitivity analyses may include using different definitions for the outcomes, comparing alternative approaches for dealing with missing data or data outliers, adjusting for baseline imbalances, performing competing risk analyses, tackling non-adherence or protocol violation, etc.[47] Findings from sensitivity analyses are considered subsidiary support for results from primary analyses. Since they do not have a confirmatory nature, it is not required to conduct multiplicity adjustments for sensitivity analyses.

What if we want to conduct multiple secondary analyses using the trial data to answer other research questions?

Because conducting clinical trials is usually costly and time-consuming, it is not uncommon to perform secondary analyses using the trial database to address other research questions. In fact, it is becoming increasingly recommended to embed trials or Studies Within A Trial (SWATs), in order to resolve uncertainties about the effects of different ways of designing, conducting, analysing and interpreting evaluations of healthcare interventions.[48] An example of SWAT can be found in a telehealth intervention study that compares different methods to improve compliance in the intervention arm in patients with chronic

obstructive pulmonary disease.[49] Besides, researchers may use the data already collected in the trial to identify factors (other than the intervention itself) associated with the primary or secondary outcomes, or they may want to focus on other outcomes (rather than the primary or secondary outcomes). For example, the PROTECT (Prophylaxis for Thromboembolism in Critical Care Trial) was conducted to evaluate the effect of dalteparin versus unfractionated heparin on proximal leg deep vein thrombosis in critically ill patients.[50] Subsequently, some studies used the data to investigate the risk factors of major bleeding,[51] thrombocytopenia[52] and all-cause death,[53] and to explore the predictors and consequences of co-enrolment of critically ill patients.[54] These secondary analyses were not predefined in the protocol of the trial because they were unrelated to the main research question.[55] Some secondary analyses reflect secondary research questions generated and logged before the trial ends by the participating international trialists, whereas other secondary analyses are not nested within original analytical plans. Such approaches may represent a cost-effective way to address new questions with trial data already collected, as long as the trial database is sufficiently complete to rigorously answer the new question. Results from secondary analyses are typically used to generate hypotheses, and their findings should be confirmed in further independent studies.[5] In addition, the questions being asked in secondary analyses are generally quite different from the primary research question(s) and thus represent different families. Therefore, there is no need to adjust for multiplicity in secondary analyses.

Statistical methods to adjust for multiplicity

There are several statistical methods that have been proposed for multiplicity adjustments. The simplest, most classical and practical method is the Bonferroni adjustment based on *P*-values. In a Bonferroni adjustment, the individual comparisons are each tested at a significance level of α/k, where k is the number of comparisons and α is the desired experiment-wise error rate. In this approach, the experiment-wise error rate for the trial remains at a level of α. The disadvantages of the Bonferroni method include low power, being overly conservative and involving irrelevant null hypotheses.[1,6,8,56] Modifications to the Bonferroni approach have been proposed, including procedures by Holm[57] and Hochberg,[58] both of which are popular and prominent multiplicity adjustment methods. Recently, many advanced multiplicity adjustment procedures have been proposed and developed. Details of advanced methods are available in a tutorial published elsewhere, in which the authors summarized the methods with examples including recycling unspent significance levels when testing hierarchical hypotheses, adapting α to the findings of previous testing and consistency requirement, graphical methods that permit repeated recycling of the significance level, and grouping hypotheses into hierarchical families of hypotheses along with recycling the significance level between those families.[59]

Multiplicity-related issues in study design, analysis plan and interpretation of results

In trials with confirmatory analyses in which issues of multiplicity cannot be avoided by design, multiplicity adjustments should always be considered and should be clearly identified in the study protocol. In cases where relevant adjustments are not carried out, a clear explanation should be provided and study results must be interpreted with caution. Multiplicity adjustment can become very challenging if several levels of multiple testing exist in a trial. For example, a trial may have multi-arm treatments, evaluate several primary outcomes, require multiple interim analyses and collect data repeatedly. In such a scenario, a simple Bonferroni adjustment would be impractical and more complex statistical procedures are needed. Therefore, we recommend that trialists seek and heed the expert advice of biostatisticians or methodologists in identifying potential multiplicity problems and determining statistical approaches for addressing them during the study design and analysis stages. A detailed statistical analysis plan clearly predefined before any treatment code is broken is essential for any clinical trial; in addition, measured and appropriate consideration and interpretation of study findings in the light of multiplicity issues are of primary importance.[5] It should be noted that multiplicity issues also impact on the type II error rate and hence on study power. For example, increases in the sample size may be required to account for multiplicity adjustments that attempt to maintain the overall type I error rate.[9,60] Such multiplicity-related issues need to be taken into account in the study design, analysis plan and result interpretation for clinical trials.

Discussion

We have presented an introduction to multiplicity adjustments in clinical trials, aiming to provide a non-technical scoping overview and to reduce the confusion around multiple testing questions. Furthermore, to facilitate understanding, we summarize and categorize the relevant conditions that are needed for multiplicity adjustment considerations in terms of trial design factors including **P**opulation, **I**ntervention/**C**omparison, **O**utcome, **T**imeframe and **A**nalysis (PICOTA) in Table 1. For instance, multiplicity adjustment questions may arise when multiple subgroup populations are analysed, in which

Table 1. Relevant conditions needed for multiplicity adjustment considerations and references in terms of trial design factors in the FAQs section

Trial design factor	Related condition needed for multiplicity adjustment consideration	Reference
Population	Multiple subgroup analyses	What if we run multiple subgroup analyses for the same outcomes?
Intervention/Comparison	Multi-arm interventions	What if we have more than two study arms?
Outcome	Multiple outcomes	What if there are multiple outcomes?
	Multiple secondary outcomes	What if we have multiple secondary outcomes?
	Repeated measurements	What if we have repeated measurements for the same outcomes?
Time frame	Multiple interim analyses	What if we conduct multiple interim analyses?
Analysis	Multiple sensitivity analyses	What if we conduct multiple sensitivity analyses for the same outcomes?
	Multiple secondary analyses for other research questions	What if we want to conduct multiple secondary analyses using the trial data to answer other research questions?

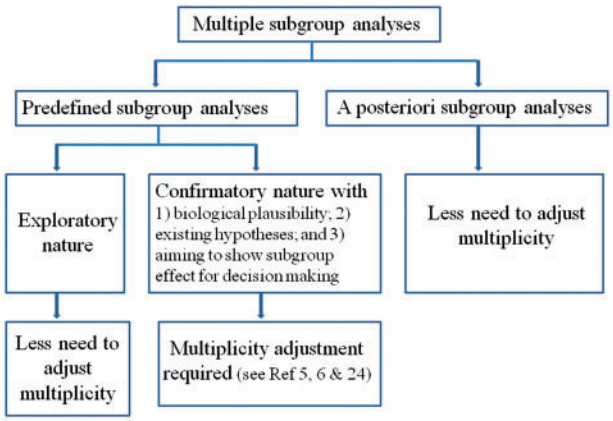

Figure 1. Flow diagram of conditions needed for multiplicity adjustment considerations and recommendations in terms of Population.

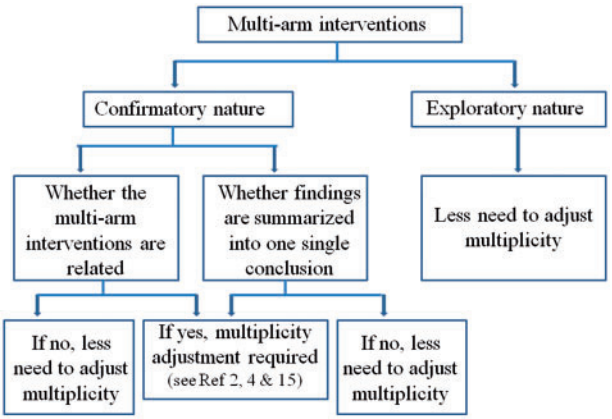

Figure 2. Flow diagram of conditions needed for multiplicity adjustment considerations and recommendations in terms of Intervention/Comparison.

more details can be referred to 'What if we run multiple subgroup analyses for the same outcomes?' in the FAQs section and in Table 1. Moreover, flow diagrams are provided in Figures 1 to 5 with key information and references according to the PICOTA framework.

In general, for trials dealing with common conditions, multiplicity adjustment considerations are required if there are no less than two primary hypotheses, unless one assumes that there is an explicit hierarchy in the multiple hypotheses. This would apply to trials generally, including factorial trials, and trials with superiority, equivalence and non-inferiority aims. Adjustments for multiplicity need to be consistent with the planned confirmatory hypotheses. Some concerns may be raised about the need for multiplicity adjustments in trials for rare diseases where it is difficult to obtain adequate sample sizes. It is important to remind ourselves that multiplicity adjustments primarily apply to confirmatory hypotheses and corresponding analyses. These decisions are made during the design stage of a

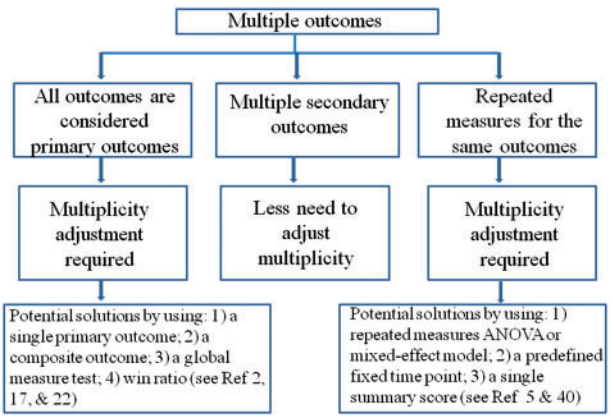

Figure 3. Flow diagram of conditions needed for multiplicity adjustment considerations and recommendations in terms of Outcome.

trial to ensure that there is adequate power to handle the necessary α adjustments. In trials for rare diseases, it is often difficult to achieve satisfactory power for a single

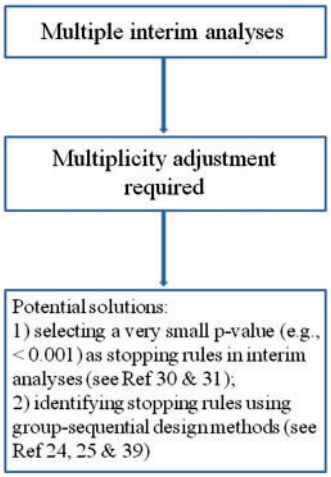

Figure 4. Flow diagram of conditions needed for multiplicity adjustment considerations and recommendations in terms of Time frame.

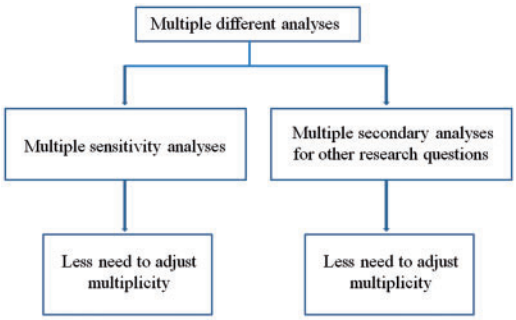

Figure 5. Flow diagram of conditions needed for multiplicity adjustment considerations and recommendations in terms of Analysis.

confirmatory hypothesis, let alone more than one.[13] As such, most trials of rare diseases either use composite outcomes as a strategy to enhance power or they are designed as exploratory investigations with no multiplicity implications. However, we recommend that the potential increased family-wise error rate should be quantified in detail and reported transparently when multiplicity adjustments are inevitably considered.

In this article, however, we did not describe detailed statistical methods for multiplicity adjustments. There are a large number of publications discussing statistical methods for multiplicity adjustments in the literature.[1,6,59–62] Herein we focus on commonly encountered questions about the need for multiplicity corrections, using a tutorial style. This may mitigate the burden of reading and understanding, especially for novice researchers.

In conclusion, multiplicity adjustments remain a challenging issue in trials, with a sizeable proportion of published trials inadequately correcting for multiplicity. Confusion around the interpretation of trial results with multiplicity issues may be reduced or avoided through careful consideration of such issues in the design stage and with a clearly pre-specified statistical analysis plan. Accordingly, sample size calculations should take into account multiplicity adjustments and the potential increase in the type II error in the study design phase.

Funding

G.L. received a Father Sean O'Sullivan Research Award, the Research Institute of St Joseph's Healthcare Hamilton, and a doctoral award from CSC. D.J.C. holds a Canada Research Chair from CIHR. This study received no specific funding.

Author contributions

All authors contributed to the study conception. G.L. drafted the first version of manuscript, and incorporated comments from other authors for revisions. All authors read and approved the final version of the manuscript. L.T. acts as the guarantor of this work.

Conflict of interest: None declared.

References

1. Proschan MA, Waclawiw MA. Practical guidelines for multiplicity adjustment in clinical trials. *Control Clin Trials* 2000;**21:** 527–39.
2. Schulz KF, Grimes DA. Multiplicity in randomised trials I:endpoints and treatments. *Lancet* 2005;**365:**1591–95.
3. Gewandter JS, Smith SM, McKeown A *et al*. Reporting of primary analyses and multiplicity adjustment in recent analgesic clinical trials: ACTTION systematic review and recommendations. *PAIN*® 2014;**155:**461–66.
4. Wason JM, Stecher L, Mander AP. Correcting for multiple-testing in multi-arm trials: is it necessary and is it done?. *Trials* 2014;**15:**364.
5. Bender R, Lange S. Adjusting for multiple testing - when and how?. *J Clin Epidemiol* 2001;**54:**343–49.
6. Cabral HJ. Multiple comparisons procedures. *Circulation* 2008;**117:**698–701.
7. Altman DG. *Practical Statistics for Medical Research*. Boca Raton, FL: CRC Press, 1990.
8. McDonald JH. *Handbook of Biological Statistics*. Baltimore, MD: Sparky House Publishing, 2009.
9. Lazzeroni LC, Ray A. The cost of large numbers of hypothesis tests on power, effect size and sample size. *Mol Psychiatry* 2012;**17:**108–14.
10. Parmar MK, Carpenter J, Sydes MR. More multiarm randomised trials of superiority are needed. *Lancet* 2014;**384**:283–84.
11. Baron G, Perrodeau E, Boutron I, Ravaud P. Reporting of analyses from randomized controlled trials with multiple arms: a systematic review. *BMC Med* 2013;**11:**84.
12. Orloff J, Douglas F, Pinheiro J *et al*. The future of drug development: advancing clinical trial design. *Nat Rev Drug Discov* 2009;**8:**949–57.
13. Koch GG, Gansky SA. Statistical considerations for multiplicity in confirmatory protocols. *Drug Inform J* 1996;**30:**523–34.

14. Fossella F, Pereira JR, von Pawel J *et al*. Randomized, multinational, phase III study of docetaxel plus platinum combinations versus vinorelbine plus cisplatin for advanced non-small-cell lung cancer: the TAX 326 study group. *J Clin Oncol* 2003;**21:** 3016–24.
15. Freidlin B, Korn EL, Gray R, Martin A. Multi-arm clinical trials of new agents: some design considerations. *Clin Cancer Res* 2008;**14:**4368–71.
16. Feise RJ. Do multiple outcome measures require p-value adjustment? *BMC Med Res Methodol* 2002;**2:**8.
17. Pocock SJ. Clinical trials with multiple outcomes: a statistical perspective on their design, analysis, and interpretation. *Control Clin Trials* 1997;**18:**530–45; discussion 46–49.
18. Turk DC, Dworkin RH, McDermott MP *et al*. Analysing multiple endpoints in clinical trials of pain treatments: IMMPACT recommendations. Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials. *Pain* 2008;**139:**485–93.
19. Montori VM, Busse JW, Permanyer-Miralda G, Ferreira I, Guyatt GH. How should clinicians interpret results reflecting the effect of an intervention on composite endpoints: should I dump this lump?. *ACP J Club* 2005;**143:**A8.
20. Ferreira-Gonzalez I, Busse JW, Heels-Ansdell D *et al*. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. *BMJ* 2007;**334:**786.
21. Ross S. Composite outcomes in randomized clinical trials: arguments for and against. *Am J Obstet Gynecol* 2007;**196:**119 e1–6.
22. Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *Eur Heart J* 2012;**33:**176–82.
23. Bakal JA, Roe MT, Ohman EM *et al*. Applying novel methods to assess clinical outcomes: insights from the TRILOGY ACS trial. *Eur Heart J* 2015;**36:**385–92.
24. Schulz KF, Grimes DA. Multiplicity in randomised trials II: subgroup and interim analyses. *Lancet* 2005;**365:**1657–61.
25. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977;**64:**191–99.
26. Pampallona S, Tsiatis AA. Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *J Stat Plann Infer* 1994;**42:**19–35.
27. Fleming TR, Harrington DP, O'Brien PC. Designs for group sequential tests. *Control Clin Trials* 1984;**5:**348–61.
28. Todd S, Whitehead A, Stallard N, Whitehead J. Interim analyses and sequential designs in phase III studies. *Br J Clin Pharmacol* 2001;**51:**394–99.
29. Stallard N, Whitehead J, Todd S, Whitehead A. Stopping rules for phase II studies. *Br J Clin Pharmacol* 2001;**51:**523–99.
30. Peto R, Pike M, 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.
31. Haybittle JL. Repeated assessment of results in clinical trials of cancer treatment. *Br J Radiol* 1971;**44:**793–97.
32. Pocock SJ, Clayton TC, Stone GW. Challenging issues in clinical trial design: Part 4 of a 4-part series on statistics for clinical trials. *J Am Coll Cardiol* 2015;**66:**2886–98.
33. Yusuf S, Sleight P, Pogue J, Bosch J, Davies R, Dagenais G. Effects of an angiotensin-converting-enzyme inhibitor, ramipril, on cardiovascular events in high-risk patients. The Heart Outcomes Prevention Evaluation Study Investigators. *N Engl J Med* 2000;**342:**145–53.
34. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979;549–56.
35. Kittelson JM, Emerson SS. A unifying family of group sequential test designs. *Biometrics* 1999;**35:**874–82.
36. Wang SK, Tsiatis AA. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* 1987;**43:** 193–99.
37. Lan KG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983;**70:**659–63.
38. Kim K, Demets DL. Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika* 1987;**74:**149–54.
39. Emerson SS, Kittelson JM, Gillen DL. Frequentist evaluation of group sequential clinical trial designs. *Stat Med* 2007;**26:** 5047–80.
40. Ludbrook J. Repeated measurements and multiple comparisons in cardiovascular research. *Cardiovasc Res* 1994;**28:** 303–11.
41. Keselman HJ, Algina J, Kowalchuk RK. The analysis of repeated measures designs: a review. *Br J Math Stat* 2001;**54:**1–20.
42. Pocock SJ, McMurray JJ, Collier TJ. Statistical controversies in reporting of clinical trials:Part 2 of a 4-part series on statistics for clinical trials. *J Am Coll Cardiol* 2015;**66:**2648–62.
43. Naggara O, Raymond J, Guilbert F, Altman D. The problem of subgroup analyses: an example from a trial on ruptured intracranial aneurysms. *Am J Neuroradiol* 2011;**32:**633–36.
44. Cook DI, Gebski VJ, Keech AC. Subgroup analysis in clinical trials. *Med J Aust* 2004;**180:**289–92.
45. Brookes ST, Whitely E, Egger M, Davey Smith G, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses: power and sample size for the interaction test. *J Clin Epidemiol* 2004;**57:**229–36.
46. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA* 1991;**266:**93–98.
47. Thabane L, Mbuagbaw L, Zhang S *et al*. A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. *BMC Med Res Methodol* 2013;**13:**92.
48. Education section - Studies Within A Trial (SWAT). *J Evid Based Med* 2012;**5:**44–45.
49. Law LM, Edirisinghe N, Wason JM. Use of an embedded, micro-randomised trial to investigate non-compliance in telehealth interventions. *Clin Trials* 2016;**13:**417–24.
50. Cook D, Meade M, Guyatt G *et al*. Dalteparin versus unfractionated heparin in critically ill patients. *N Engl J Med* 2011; **364:**1305–14.
51. Lauzier F, Arnold DM, Rabbat C *et al*. Risk factors and impact of major bleeding in critically ill patients receiving heparin thromboprophylaxis. *Intensive Care Med* 2013;**39:** 2135–43.
52. Williamson DR, Albert M, Heels-Ansdell D *et al*. Thrombocytopenia in critically ill patients receiving thromboprophylaxis:frequency, risk factors, and outcomes. *Chest* 2013;**144:**1207–15.
53. Li G, Thabane L, Cook DJ *et al*. Risk factors for and prediction of mortality in critically ill medical-surgical patients

receiving heparin thromboprophylaxis. *Ann Intensive Care* 2016;**6:**18.
54. Cook D, McDonald E, Smith O *et al*. Co-enrollment of critically ill patients into multiple studies: patterns, predictors and consequences. *Crit Care* 2013;**17:**R1.
55. Cook D, Meade M, Guyatt G *et al*. PROphylaxis for ThromboEmbolism in Critical Care Trial protocol and analysis plan. *J Crit Care* 2011;**26:**223 e1–9.
56. Perneger TV. What's wrong with Bonferroni adjustments. *BMJ* 1998;**316:**1236.
57. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat* 1979;**6:**65–70.
58. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 1988;**75:**800–02.
59. Alosh M, Bretz F, Huque M. Advanced multiplicity adjustment methods in clinical trials. *Stat med* 2014;**33:**693–713.
60. Shaffer JP. Multiple hypothesis testing. *Ann Rev Psychol* 1995;**46:**561–84.
61. Westfall PH, Tobias RD, Wolfinger RD. *Multiple Comparisons and Multiple Tests Using SAS*. Cary, NC: SAS Institute, 2011.
62. Bretz F, Hothorn T, Westfall P. *Multiple Comparisons Using R*. Boca Raton, FL: CRC Press, 2010.