Trials

COMMENTARY Open Access

Multiple secondary outcome analyses: precise interpretation is important

Richard A. Parker* and Christopher J. Weir

Abstract

Analysis of multiple secondary outcomes in a clinical trial leads to an increased probability of at least one false significant result among all secondary outcomes studied. In this paper, we question the notion that that if no multiplicity adjustment has been applied to multiple secondary outcome analyses in a clinical trial, then they must necessarily be regarded as exploratory. Instead, we argue that if individual secondary outcome results are interpreted carefully and precisely, there is no need to downgrade our interpretation to exploratory. This is because the probability of a false significant result for each comparison, the per-comparison wise error rate, does not increase with multiple testing. Strong effects on secondary outcomes should always be taken seriously and must not be dismissed purely on the basis of multiplicity concerns.

Keywords: Multiple testing, Secondary outcomes, Family-wise error rate, Per-comparison wise error rate, Type-I error, Multiplicity

Background

Clinical trials commonly designate one or more outcomes as primary and the rest as secondary or exploratory. Multiple testing of secondary outcomes leads to an increase in the family-wise error rate (FWER): the probability of at least one false significant result (type I error) among all secondary outcomes studied. This is why research studies often adjust for multiplicity when performing secondary outcome analyses. However, this is not always necessary. In particular, if all secondary outcome results are required to be statistically significant to conclude an overall treatment effect, then the type I error at the level of the individual tests is not relevant, and multiple testing adjustment is not appropriate [1, 2].

Multiple testing adjustment is usually achieved using a multiple testing correction, such as Bonferroni, to achieve overall type I error rate control [3–6]; but there also appears to be a growing trend for pre-specified hierarchical testing structures, whereby the secondary outcomes are tested in a pre-specified order [7–9] or by using a more complex graphical adjustment scheme [10–12]. Alternatively, researchers frequently adopt an exploratory interpretation. For example, in a clinical trial by Cao et al. [13] of a COVID-19 treatment conducted in 2020, no correction for multiplicity was applied, and the authors indicated that "the [confidence] intervals should not be used to infer definitive treatment effects for secondary outcomes" [13]. Another example is the RICH trial publication [14], where the authors write "Statistical analyses of secondary end points were not adjusted for multiplicity. Therefore, because of the potential for type I error, findings should be interpreted as exploratory." [14] Both of these examples are consistent with the widely reported belief that if no multiplicity adjustment has been applied to multiple secondary outcomes in a clinical trial, then they must necessarily be regarded as exploratory [15–17]. In this article, we outline some counter-arguments to this view before suggesting a careful and valid way to interpret multiple secondary outcomes that neither requires an exploratory

* Correspondence: Richard.Parker@ed.ac.uk
Edinburgh Clinical Trials Unit, Usher Institute, University of Edinburgh, Edinburgh, UK

© The Author(s). 2022 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

interpretation of secondary outcomes nor requires formal multiple testing adjustment.

Issues with the "exploratory interpretation requirement" view

One interpretational issue with necessitating an exploratory interpretation of secondary outcomes in the absence of multiplicity adjustment is that pre-defined secondary endpoints will be classified in the same category as outcomes of lower importance in a trial, noting that many trials will pre-specify a list of secondary endpoints and a separate list of exploratory (or tertiary) endpoints. By considering all non-primary outcomes as exploratory, this would preclude the need to distinguish between the two. Such an approach, however, fails to recognize the many situations where secondary outcomes are critical to the overall interpretation of a clinical trial, for example in the case where the individual components of a composite primary outcome are included among the secondary outcomes. Secondary endpoint results will often feed directly into the overall interpretation of a clinical trial and facilitate understanding of the scope of any potential intervention effect. In contrast, tertiary endpoints are likely to be more exploratory, novel, hypothesis-generating, or mechanistic. For example, tertiary endpoints could be used to explore disease processes or the mechanisms by which an intervention is effective. By themselves, these tertiary endpoints are unlikely to modify the key conclusions in a clinical trial, but they may still be of interest for guiding future research. It is therefore important to retain the distinction between the two to facilitate appropriate trial interpretation.

Another problem is that if a secondary variable shows a clinically important effect or is highly significant (for example, with $P < 0.0001$), as one of the secondary outcomes in Baggot et al. [17] was found to be, does this mean that it is *only* permissible to interpret this result as exploratory if no adjustment has been made for multiple testing? Based on Gao et al., the probability of a false positive conclusion (type I error) conditional on this *p*-value will be no greater than 0.018 [18]. Therefore, given that we will have confidence that such a significant result is real and replicable, is it really fair to downgrade this finding to "exploratory"? Indeed, there is the danger that strong adverse effects on safety outcomes may not be taken seriously if they are among multiple outcomes.

To highlight the reality of this danger, consider the large COMPASS trial reported in 2019 [19], which randomized over 17,000 patients to receive proton pump inhibitors (PPIs) [19]. The trial authors reported a statistically significant increased risk of enteric infections in those allocated to a PPI (odds ratio 1.33, 95% confidence interval 1.01 to 1.75) [19]. However, in their trial publication, the authors stated that the "data in the current randomized trial were not adjusted for multiple testing, so this result should be interpreted with caution," even though "enteric infection" was a pre-specified safety outcome [19, 20]. Although caution may indeed have been merited on the basis of the relatively high *p*-value (0.04), modest odds ratio, and wide confidence interval; we would question whether any safety signals should be treated any less seriously simply because they were one among many different safety outcomes.

A third problem is that it may lead researchers and other stakeholders to implicitly link the value of secondary outcome results with the number of secondary outcomes reported, such that their value may depreciate as the number of outcomes reported increases.

The "per-comparison-wise error rate" approach

The relationship between the number of tests performed and the overall type I error rate is well known; what is much less well appreciated is that the *probability* of a false significant result for *each* comparison of an outcome between treatment groups, the "per-comparison-wise error rate" (PCWER), does not increase with multiple testing [2, 21, 22]. That means that if we adopt a precise, focused interpretation of the individual results, then there is no need to either apply a multiplicity adjustment or downgrade our interpretation to "exploratory." Indeed, we would argue that strong interpretations could be made if secondary outcomes are interpreted very precisely and carefully according to the specific variables used in the analysis. This means, for example, instead of vaguely indicating that the "intervention has an effect," we more specifically state that the "intervention has a specific effect on systolic blood pressure at 12 weeks."

Furthermore, to take the example considered earlier of a clinical trial in COVID-19, Cao et al. write: "28-day mortality was numerically lower in the lopinavir–ritonavir group than in the standard-care group" [13]. In this case, the authors are making a specific statement about treatment efficacy for the secondary outcome of mortality at 28 days. The probability of a false significant result is only inflated if the authors were claiming generic treatment efficacy from this one outcome; but since the statement is specific to the 28-day mortality outcome, then there is no need to downgrade the interpretation to "exploratory."

Similar arguments have been made by Rubin [1], who reasons that individual testing does not require multiplicity adjustment if we are making "specific inference" relating to individual null hypotheses [1].

Precise interpretation involves interpreting outcomes in a way that refers to all of their distinguishing features (e.g., time point, type of outcome, intervention) as appropriate, so that we can differentiate clearly between underlying individual hypothesis tests.

Advantages of the precise interpretation

Firstly, an advantage of using the precise interpretation is that it prevents pre-specified secondary outcomes, especially safety outcomes, from being undervalued or dismissed as unimportant. As noted previously, strong effects on secondary outcomes are highly likely to be real and replicable [18] and therefore must be taken seriously regardless of the number of comparisons performed.

Secondly, it promotes a careful interpretation of secondary outcomes that is not tied to the number of multiple tests performed. After all, it does not make sense for the total number of secondary outcomes to have a bearing on how the finding for each individual outcome is interpreted [23].

Thirdly, it reduces the potential for selective reporting bias, whereby in fear of the negative consequences or perceptions of multiple testing, authors only select the most interesting secondary outcome results for formal reporting [24].

Note of caution

A note of caution is needed at this point however. If a main trial result is *not* significant, and a secondary outcome is simply the primary outcome measured at a different time point, then care is needed that this does not becomes a way of inferring an overall treatment effect by the “back door.” Success of a treatment on a secondary outcome should not in general overcome a lack of benefit found on the primary, even if that secondary outcome were pre-specified.

More generally, if a set of secondary outcomes consist of the same outcome measured at different time points, and only one of these outcomes is statistically significant, precise interpretation requires us to take great care not to conclude that an intervention is effective on the basis of a single result without reference to the findings at the other time points. Such care may also be needed if different secondary outcomes are expected to be highly correlated with each other. In this case, interpretation of a single outcome cannot be divorced from the other outcomes that are highly correlated. This is one reason why it is crucial to transparently report *all* related secondary outcome results and statistical tests within the same study report or publication and ensure that one secondary outcome result is not unduly promoted over and above other secondary outcomes that are strongly related [25].

In contrast, for *distinct* secondary outcome comparisons, it is perfectly natural to make individual interpretations for each comparison and that this is reflected by control of the PCWER rate rather than control of the FWER. As we have argued elsewhere in the context of multi-arm trials, it does not make sense to control the FWER rate in this situation when our interpretation is at the per-comparison level [21].

Conclusions

Multiple secondary outcomes are by definition subsidiary to primary outcomes, but this does not mean that they must necessarily be downgraded to the level of exploratory in the absence of multiplicity adjustment. Indeed, if individual secondary outcome results are interpreted precisely, then the number of tests performed is irrelevant because the per-comparison-wise error rate is not increased. What we are advocating is a careful and precise interpretation of secondary outcome results. Strong effects on secondary outcomes should always be taken seriously and must not be dismissed purely on the basis of multiplicity concerns.

Abbreviations

PCWER: Per comparison-wise error rate; FWER: Family-wise error rate

Acknowledgements

Not applicable.

Authors’ contributions

RP had the initial idea for the paper, wrote the first draft, and was involved in revising the paper. CW was involved with drafting the paper and revising it critically for important intellectual content. The authors read and approved the final manuscript.

Funding

None.

Availability of data and materials

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 27 August 2021 Accepted: 21 December 2021
Published online: 10 January 2022

References

1. Rubin M. When to adjust alpha during multiple testing: a consideration of disjunction, conjunction, and individual testing. Synthese. 2021;6:1–32.
2. Dmitrienko A, Bretz F, Westfall PH, Troendle J, Wiens BL, Tamhane AC, et al. Multiple testing methodology. In: Dmitrienko A, Tamhane AC, Bretz F, editors. Multiple testing problems in pharmaceutical statistics. 1st ed. Boca Raton, Florida, U.S.A: Chapman & Hall/CRC Biostatistics Series; 2010. p. 35–41.
3. Schober P, Vetter TR. Adjustments for multiple testing in medical research. Anesth Analg. 2020;130(1):99.
4. Turan A, Duncan A, Leung S, Karimi N, Fang J, Mao G, et al. Dexmedetomidine for reduction of atrial fibrillation and delirium after cardiac surgery (DECADE): a randomised placebo-controlled trial. Lancet. 2020;396(10245):177–85.

5. Lee N, Gao Y, Collins SL, Mårtensson LB, Randall W, Rowe TM, et al. Caesarean delivery rates and analgesia effectiveness following injections of sterile water for back pain in labour: a multicentre, randomised placebo controlled trial. EClinicalMedicine. 2020 Aug;1(25):100447.
6. Parker R, (Academic). Introduction to multiple testing and methods for error rate control [Video]: SAGE Research Methods Video: Medicine and Health; 2021. https://doi.org/10.4135/9781529778205.
7. Furtado RH, Berwanger O, Fonseca HA, Corrêa TD, Ferraz LR, Lapa MG, et al. Azithromycin in addition to standard of care versus standard of care alone in the treatment of patients admitted to the hospital with severe COVID-19 in Brazil (COALITION II): a randomised clinical trial. Lancet. 2020;396(10256): 959–67.
8. Paz-Ares L, Ciuleanu TE, Cobo M, Schenker M, Zurawski B, Menezes J, et al. First-line nivolumab plus ipilimumab combined with two cycles of chemotherapy in patients with non-small-cell lung cancer (CheckMate 9LA): an international, randomised, open-label, phase 3 trial. Lancet Oncol. 2021; 22(2):198–211.
9. Charytan C, Manllo-Karim R, Martin ER, Steer D, Bernardo M, Dua SL, et al. A randomized trial of roxadustat in anemia of kidney failure: SIERRAS study. Kidney Int Rep. 2021; Apr 17.
10. Marconi VC, Ramanan AV, de Bono S, Kartman CE, Krishnan V, Liao R, et al. Efficacy and safety of baricitinib for the treatment of hospitalised adults with COVID-19 (COV-BARRIER): a randomised, double-blind, parallel-group, placebo-controlled phase 3 trial. Lancet Respir Med. 2021;1.
11. Deodhar A, van der Heijde D, Gensler LS, Kim TH, Maksymowych WP, Østergaard M, et al. Ixekizumab for patients with non-radiographic axial spondyloarthritis (COAST-X): a randomised, placebo-controlled trial. Lancet. 2020;395(10217):53–64.
12. Ashina M, Reuter U, Smith T, Krikke-Workel J, Klise SR, Bragg S, et al. Randomized, controlled trial of lasmiditan over four migraine attacks: findings from the CENTURION study. Cephalalgia. 2021;41(3):294–304.
13. Cao B, Wang Y, Wen D, Liu W, Wang J, Fan G, et al. A trial of lopinavir–ritonavir in adults hospitalized with severe COVID-19. New Engl J Med. 2020; Mar 18.].
14. Zarbock A, Küllmar M, Kindgen-Milles D, Wempe C, Gerss J, Brandenburger T, et al. Effect of regional citrate anticoagulation vs systemic heparin anticoagulation during continuous kidney replacement therapy on dialysis filter life span and mortality among critically ill patients with acute kidney injury: a randomized clinical trial. JAMA. 2020;324(16):1629–39.
15. Li G, Taljaard M, Van den Heuvel ER, Levine MA, Cook DJ, Wells GA, et al. An introduction to multiplicity issues in clinical trials: the what, why, when and how. Int J Epidemiol. 2017;46(2):746–55.
16. Khan MS, Khan MS, Ansari ZN, Siddiqi TJ, Khan SU, Riaz IB, et al. Prevalence of multiplicity and appropriate adjustments among cardiovascular randomized clinical trials published in major medical journals. JAMA Netw Open. 2020;3(4):e203082.
17. Baggott C, Reddel HK, Hardy J, Sparks J, Holliday M, Corin A, et al. Patient preferences for symptom-driven or regular preventer treatment in mild to moderate asthma: findings from the PRACTICAL study, a randomised clinical trial. Eur Respir J. 2020;55(4):1902073.
18. Gao J. *P*-values – a chronic conundrum. BMC Med Res Methodol. 2020; 2020(167). https://doi.org/10.1186/s12874-020-01051-6.
19. Moayyedi P, Eikelboom JW, Bosch J, Connolly SJ, Dyal L, Shestakovska O, et al. Safety of proton pump inhibitors based on a large, multi-year, randomized trial of patients receiving rivaroxaban or aspirin. Gastroenterology. 2019;157(3):682–91.
20. Bosch J, Eikelboom JW, Connolly SJ, Bruns NC, Lanius V, Yuan F, et al. Rationale, design and baseline characteristics of participants in the cardiovascular outcomes for people using anticoagulation strategies (COMPASS) trial. Can J Cardiol. 2017;33(8):1027–35.
21. Parker RA, Weir CJ. Non-adjustment for multiple testing in multi-arm trials of distinct treatments: rationale and justification. Clin Trials. 2020. https://doi.org/10.1177/1740774520941419.
22. Bretz F, Koenig F. Commentary on Parker and Weir. Clin Trials. 2020. https://doi.org/10.1177/1740774520941420.
23. Perneger TV. What's wrong with Bonferroni adjustments. BMJ. 1998;315: 1236–8.
24. Althouse AD. Adjust for multiple comparisons? It's not that simple. Ann Thoracic Surg. 2016;101(5):1644–5.
25. Patel CJ, Ioannidis JPA. Placing epidemiological results in the context of multiplicity and typical correlations of exposures. J Epidemiol Commun Heal. 2014;68:1096–100.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC