Handling multiple primary outcomes in randomised controlled trials: An overview

Rumana Z Omar, Victoria Vickerstaff and Gareth Ambler

Department of Statistical Science

University College London

UK

Introduction

Randomised controlled trials (RCT) are the most rigorous way to investigate the effectiveness of a new intervention on a health outcome. When investigating the effect of an intervention in a trial, the conventional approach is to select a single primary outcome which adequately represents the health condition of interest. Limiting the primary comparison to a single outcome is advised by the regulatory bodies. However it may not be possible to characterise many health conditions using a single outcome when evaluating the effect of an intervention. For example, trials in mental health disorders, stroke and chronic obstructive pulmonary disease may need more than one primary outcome to be considered to provide a comprehensive understanding of the effects of the intervention on the health condition of interest.

Often multiple statistical tests are performed when analysing multiple outcomes in trials. When multiple statistical tests are performed, there is a chance that a statistically significant result will be observed due to chance when actually no effect is present. This is known as a '*type I error*'. As the number of statistical tests performed on the same dataset increases, the probability of a type I error increases. It may be important to control for multiplicity in confirmatory phase III where the goal of the trial is to confirm the effect of an intervention. In contrast, when performing an early phase drug trial one is usually interested to explore the effects of different drug doses and therefore addressing multiplicity is less important for these types of trial designs.

Several methods have been proposed in the literature to address the problem of multiplicity. For the practitioner, it is often unclear which (if any) of the proposed methods should be used to account for multiplicity whilst ensuring that the analysis remains efficient. It is also important that the methodology is accessible to clinicians who need to interpret the findings. Inevitably, it is not always possible to measure the required outcomes for all participants and thus missing outcome data is a common problem for RCTs. In fact, the majority of the published trials report missing data which may reduce the power. If the study does not have

sufficient power, then true intervention effects may not be detected. It is important to choose an adjustment method for multiplicity that performs well in the presence of missing data.

Methods of analysis for trials with multiple outcomes

There are several methods available for the analysis of multiple outcomes in a trial. A commonly used method is to analyse each outcome separately within a univariate framework. This approach is appealing due its simplicity however a limitation is that it does not account for the possible correlation between the outcomes, which may result in a loss of efficiency in the analysis leading to less power to detect the intervention effects. Additionally, observations with missing data may be omitted from parts of the analyses.

Multiple imputation may be used to impute missing data prior to the analysis and the correlation between the outcomes can be accounted for by the imputation model. More advanced techniques such as multivariate models have been introduced that enable multiple outcomes to be analysed simultaneously by taking into account the correlations between them. A multivariate approach is perhaps more suitable to characterise a particular health condition which cannot be represented adequately by a single primary outcome. It may also increase the efficiency in the estimation of the intervention effect by accounting for the correlation between the outcomes. Examples of multivariate methods include the multivariate multilevel model and multivariate regression. These multivariate methods have been used to analyse examination results in schools and crime trends, however, their application in trials has been limited. Multivariate methods may also be more efficient when some of the outcomes have missing values. Multiple primary outcomes may have the same data type, for example, several continuous outcomes may be measured concurrently to evaluate the effect of cognitive behavioural therapy for patients with a depressive disorder. In this case, researchers may wish to examine both the cognitive and behavioural components. For some trials, the data on a mixture of outcome types may be collected. For example, a mixture of continuous and binary outcomes may be collected to evaluate the effect of an antipsychotic drug for people with schizophrenia and researchers could examine both quality of life (a continuous outcome) and whether the participant has a symptom relapse or not (a binary outcome). Multivariate models can typically handle multiple continuous outcomes, binary outcomes or a mixture of both. Multivariate multilevel (MM) models can handle non-overlapping missingness across outcomes (when values are missing for some of the outcomes). Therefore it does not require the number of observations to be balanced across outcomes. This has implications for the handling of missing data, because if some of the outcome data are missing, more efficient estimates of the

intervention effect may be estimated by making use of the correlations between the outcome measures.

Regardless of whether a univariate or multivariate framework is used for the analysis, some form of adjustment for multiplicity would need to be applied to preserve the familywise error rate.

Sample size and power

A sample size calculation is usually performed for trials to ensure that sufficient participants are recruited to achieve a desired level of power. In the context of multiple outcomes, the power of the study can be defined in a number of ways depending on the clinical objective of the trial, for example: i) 'disjunctive power', ii) 'conjunctive power' or iii) 'marginal power'. The disjunctive power (or minimal power) is the probability of finding at least one true intervention effect across all of the outcomes. The conjunctive power (or maximal power) is the probability of finding a true intervention effect on all outcomes. The marginal (or individual) power is the probability of finding a true intervention effect on a particular outcome and is calculated separately for each outcome. When the clinical objective of a trial is to detect a statistically significant intervention effect for at least one of the outcomes the disjunctive power and marginal power are recommended whereas the conjunctive power is recommended when the clinical objective is to detect a statistically significant intervention effect for all the outcomes. In this paper, we focus on the former clinical objective and therefore we focus on disjunctive and marginal power.

An approach often used in trials is to calculate the sample size separately for each of the primary outcomes by applying a Bonferroni correction to adjust the significance level. The largest value of the sample size is then considered as the final sample size for the trial. However, the power requirements of a trial should match the clinical objective which needs to be pre-specified when designing the study and the sample size calculation should be performed accordingly.

Literature review

We conducted a review of 209 RCTs in the field of neurology and psychiatry published in high impact journals. Our review found that multiple primary outcomes were commonly used but often inadequately handled in trials. The most commonly used method to analyse multiple primary outcomes in published randomised trials was to analyse them separately within the univariate framework. Only a small number of trials accounted for multiplicity and Bonferroni's adjustment was the most commonly used method.

Simulation studies

We conducted simulation studies to investigate the disjunctive power, marginal power and familywise error rate (FWER) obtained using different adjustment methods for multiplicity. These were the Bonferroni, Holm, Hochberg, Dubey/Armitage-Parmar and Stepdown-minP adjustment methods. Different simulation scenarios were constructed by varying the number of outcomes, degree of correlation between the outcomes, intervention effect sizes and proportion of missing data. Simulation studies were also used to investigate the disjunctive power and FWER obtained when using the MM model in comparison to those obtained when analysing outcomes separately using univariate models. In the presence of missing data, multiple imputation is used to impute missing outcomes when analysing the outcomes separately (MI+UV). Additional simulation scenarios for this comparison included varying the types of outcomes: 1) all continuous outcomes; 2) all binary outcomes; and 3) a mix of continuous and binary outcomes. We also performed simulations in which the missing data is missing not at random (MNAR) to investigate the bias in the estimated treatment effects when using the MM model and the MI+UV compared to analysing outcomes separately using UV.

Findings:

Our simulation studies suggest that the Bonferroni adjustment should be used for the sample size calculation of RCTs with multiple primary outcomes since it preserves the FWER with little loss of power. For the analysis, when outcomes have missing values, we suggest based on our simulation studies that either the Hochberg or Hommel methods are used to account for multiplicity provided that the distributional assumptions are met. These two methods provide the highest power. The sample size requirement to achieve the desired disjunctive power may be substantially smaller than that required to achieve the desired marginal power. The choice between whether to specify a disjunctive or marginal power should depend on the clinical objective and this should be pre-specified. With regards to multiplicity arising from multiple outcomes, CONSORT states that “authors should exercise special care when evaluating the results of trials with multiple comparisons”. We recommend that the chosen method to maintain the FWER at the desired level is described and justification for the choice provided in the statistical analysis plan for trials.

The simulation studies comparing the multivariate and univariate methods show that when MM is used with a Holm adjustment, the FWER fluctuates around 5%. In terms of disjunctive power, the MM performs better than using UV when the outcomes are correlated and in the presence of missing data. There was a notable increase in power when the correlation between the outcomes exceeded 0.4. MM model offers a computational advantage to multiple imputation as the MM enables the analysis to be performed in just one step. In

contrast, multiple imputation requires three steps: specifying the imputation model and performing the imputation, fitting the analysis model to each imputed datasets and then combining the results across the imputed datasets.

The simulation studies investigating MNAR data suggest that when MI+UV and MM are used, no gains in terms of bias may be made if there is no correlation between the outcome measures. There were gains in terms of bias for both the MI+UV and MM methods when the outcomes are highly correlated and in the presence of high levels of missing data. There was a notable decrease in bias when the correlation exceeds 0.4. The MM approach appeared to outperform the MI+UV in the more extreme cases of high levels of missing data. However, neither approach was able to remove the bias completely. As a consequence, any inferences and conclusions made within the trial setting would need to be tested with sensitivity analyses under the alternative assumption that the missing data are MNAR.

References

1. Agusti, A. & Vestbo, J. 2011. Current controversies and future perspectives in chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine,* 184, 507-13.
2. Andromachi Tseloni & Christina Zarafonitou**.** Fear of Crime and Victimization: A Multivariate Multilevel Analysis of Competing Measurements. European Journal of Criminology, 2008. https://doi.org/10.1177/1477370808095123.
3. Bell ML, Fiero M, Horton NJ, Hsu C-H: Handling missing data in RCTs; a review of the top Medical journals. *BMC Medical Research Methodology* 2014, **14**(1):118.
4. Bender R, Lange S: Adjusting for multiple testing—when and how? *Journal of clinical Epidemiology* 2001, **54**(4):343-349.
5. Bretz F, Posch M, Glimm E, Klinglmueller F, Maurer W, Rohmeyer K: Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests. *Biometrical Journal* 2011, **53**(6):894-913.
6. Bretz F, Hothorn T, Westfall P: Multiple comparisons using R: CRC Press; 2010.
7. Blakesley RE, Mazumdar S, Dew MA, Houck PR, Tang G, Reynolds III CF, Butters MA: Comparisons of methods for multiple hypothesis testing in neuropsychological research. *Neuropsychology* 2009, **23**(2):255.
8. Campbell AN, Nunes EV, Matthews AG, Stitzer M, Miele GM, Polsky D, Turrigiano E, Walters S,McClure EA, Kyle TL: Internet-delivered treatment for substance abuse: a multisite randomized controlled trial. *American Journal of Psychiatry* 2014, **171**(6):683-690.
9. Capizzi T, Zhang J: Testing the hypothesis that matters for multiple primary endpoints. *Drug information journal* 1996, **30**(4):949-956.
10. Chow S-C, Shao J, Wang H, Lokhnygina Y: Sample size calculations in clinical research: Chapman and Hall/CRC; 2017.
11. De Los Reyes A, Kundey SMA, Wang M: The end of the primary outcome measure: A Research agenda for constructing its replacement. *Clinical Psychology Review* 2011, **31**(5):829-838.
12. Dmitrienko A, Tamhane AC, Bretz F: Multiple testing problems in pharmaceutical statistics: CRC Press; 2009.
13. Dmitrienko A, D'Agostino R: Traditional multiplicity adjustment methods in clinical trials. S*tatistics in Medicine* 2013, **32**(29):5172-5218.
14. European Medical Agency: Guideline on multiplicity issues in clinical trials. In.; 2017.
15. Food, Administration D: Multiple Endpoints in Clinical Trials Guidance for Industry. *FDA Issues Draft Guidance* 2017.
16. Ge Y, Dudoit S, Speed TP: Resampling-based multiple testing for microarray data analysis. *Test* 2003, **12**(1):1-77.
17. Gelman A, Hill J, Yajima M: Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness* 2012, **5**(2):189-211.
18. GOLDSTEIN, H. 2011. *Multilevel statistical models*, Wiley. com.
19. Hassiotis A[1], Robotham D, Canagasabey A, Romeo R, Langridge D, Blizard R, S Murad , King M. Randomized, single-blind, controlled trial of a specialist behavior therapy team for challenging behavior in adults with intellectual disabilities. Am J Psychiatry. 2009 Nov;166(11):1278-85. doi:10.1176/appi.ajp.2009.08111747. Epub 2009 Aug 17.
20. Holm S: A simple sequentially rejective multiple test procedure. *Scandinavian journal of Statistics* 1979:65-70.
21. Hochberg Y: A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 1988, **75**(4):800-802.
22. Hommel G: A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* 1988, **75**(2):383-386.

23. Lafaye de Micheaux P, Liquet B, Marque S, Riou J: Power and sample size determination in clinical trials with multiple primary continuous correlated endpoints. *Journal of biopharmaceutical statistics* 2014, **24**(2):378-397.
24. Li D, Dye TD: Power and stability properties of resampling-based multiple testing procedures with applications to gene oncology studies. *Computational and mathematical methods inmedicine* 2013.
25. MAYO, N. E. & SCOTT, S. 2011. Evaluating a complex intervention with a single outcome may not be a good idea: an example from a randomised trial of stroke case management. *Age and ageing,* 40**,**718-724.
26. Middleton S, McElduff P, Ward J, Grimshaw JM, Dale S, D'Este C, Drury P, Griffiths R,Cheung NW, Quinn C: Implementation of evidence-based treatment protocols to manage fever, hyperglycaemia, and swallowing dysfunction in acute stroke (QASC): a cluster randomised controlled trial. *The Lancet* 2011, **378**(9804):1699-1706.
27. PHILLIPS, A. & HAUDIQUET, V. 2003. ICH E9 guideline 'Statistical principles for clinical trials': a case study. *Statistics in Medicine,* 22**,** 1-11.
28. PRODUCTS, C. F. P. M. 2002. Points to consider on multiplicity issues in clinical trials. *London: TheEuropean Agency for the Evaluation of Medicinal Products (EMEA)*.
29. POCOCK, S. J. 1997b. Clinical trials with multiple outcomes: a statistical perspective on their design, analysis, and interpretation. *Controlled clinical trials,* 18**,** 530-545.
30. Reitmeir P, Wassmer G: Resampling-based methods for the analysis of multiple endpoints clinical trials. *Statistics in Medicine* 1999, **18**(24):3453-3462.
31. SCHULZ, K. F., ALTMAN, D. G. & MOHER, D. 2010. CONSORT 2010 statement:updated guidelines for reporting parallel group randomised trials. *BMC medicine,* 8**,** 18.
32. Senn S, Bretz F: Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics* 2007, **6**(3):161-170.
33. Shaffer JP: Multiple hypothesis testing. *Annual review of psychology* 1995, **46**(1):561-584.
34. Sugimoto T, Sozu T, Hamasaki T: A convenient formula for sample size calculations in clinical trials with multiple co-primary continuous endpoints. *Pharmaceutical Statistics* 2012, 11(2):118-128.
35. Suresh K, Chandrashekara S: Sample size estimation and power analysis for clinical Research studies. *Journal of human reproductive sciences* 2012, **5**(1):7.
36. SNIJDERS, T. & BOSKER, R. J. 2012. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling, second edition. ,* London Sage Publishers.
37. Teixeira-Pinto A, Siddique J, Gibbons R, Normand S-L: Statistical approaches to modeling multiple outcomes in psychiatric studies. *Psychiatric annals* 2009, **39**(7):729.
38. Vickerstaff V, Ambler G, King M, Nazareth I, Omar RZ: Are multiple primary outcomes analysed appropriately in randomised controlled trials? A review. *Contemporary clinical trials* 2015, **45**:8-12.
39. Wright SP: Adjusted p-values for simultaneous inference. *Biometrics* 1992:1005-1013.
40. Westfall PH, Young SS: Resampling-based multiple testing: Examples and methods for p-value adjustment, vol. 279: John Wiley & Sons; 1993.Warner RM: Applied statistics: From bivariate through multivariate techniques: Sage;2008.