

Vickerstaff, V; Ambler, G; King, M; Nazareth, I; Omar, RZ; (2015) Are multiple primary outcomes analysed appropriately in randomised controlled trials? A review. **Contemporary Clinical Trials** [10.1016/j.cct.2015.07.016](https://doi.org/10.1016/j.cct.2015.07.016). (In press). Downloaded from UCL Discovery: <http://discovery.ucl.ac.uk/1470150>

ARTICLE

Are multiple primary outcomes analysed appropriately in randomised controlled trials? A review.

Victoria Vickerstaff^{1 2 3}, *Gareth Ambler*², *Michael King*¹, *Irwin Nazareth*³,
*Rumana Z Omar*²

¹*UCL Division of Psychiatry, University College London 6th Floor, Maple House, 149 Tottenham Court Road, London, W1T 7NF*

²*Department of Statistical Science, University College London, Gower Street, London, WC1E 6BT*

³*The Research Department of Primary Care and Population Health, University College London, Rowland Hill Street, London, NW3 2PF, UK*

**Corresponding author:*

¹*Marie Curie Palliative Care Research Department,
UCL Division of Psychiatry
University College Medical School,
6th Floor Maple House,
149 Tottenham Court Road, London
W1T 7NF Email: v.vickerstaff@ucl.ac.uk
Telephone: 020 7679 9721*

Abstract:

Objectives To review how multiple primary outcomes are currently considered in the analysis of randomised controlled trials. We briefly describe the methods available to safeguard the inferences and to raise awareness of the potential problems caused by multiple outcomes.

Methods/design We reviewed randomised controlled trials (RCTs) in neurology and psychiatry disease areas, as these frequently analyse multiple outcomes. We reviewed all published RCTs from July 2011 to June 2014 inclusive in the following high impact journals: The New England Journal of Medicine, The Lancet, The American Journal of Psychiatry, JAMA Psychiatry, The Lancet Neurology and Neurology. We examined the information presented in the abstract and the methods used for sample size calculation and statistical analysis. We recorded the number of primary outcomes, the methods used to account for multiple primary outcomes, the number of outcomes discussed in the abstract and the number of outcomes used in the sample size calculation.

Results Of the 209 RCTs that we identified, 60 (29%) analysed multiple primary outcomes. Of these, 45 (75%) did not adjust for multiplicity in their analyses. Had multiplicity been

addressed, some of the trial conclusions would have changed. Of the 15 (25%) trials which accounted for multiplicity, Bonferroni's correction was the most commonly used method.

Conclusions Our review shows that trials with multiple primary outcomes are common. However, appropriate steps are not usually taken in most of the analyses to safeguard the inferences against multiplicity. Authors should state their chosen primary outcomes clearly and justify their methods of analysis.

Keywords: Multiplicity; multiple outcomes; clinical trials; neurology; psychiatry

Introduction

In a randomised controlled trial, a single outcome may be insufficient to describe all the effects of an intervention on a complex disease. Multiple health outcomes may need to be investigated to assess all the relevant aspects of the disorder (1). These health outcomes are often correlated. Neurology (2) and psychiatry (3) are two disease areas where multiple primary outcomes are particularly needed to evaluate a health intervention, for example when evaluating depression (4), stroke (5) and long-term mental health conditions (1). To evaluate the effect of the intervention on multiple primary outcomes in a trial, each outcome could be analysed separately (6). However, if the multiple outcomes are not accounted for in the statistical analysis appropriately, the probability of obtaining statistically significant results by chance may increase. The probability of finding at least one false significant result is called the familywise error rate (FWER)(7). The FWER should be maintained at an acceptable level, usually 5.0% (8).

If each outcome is analysed separately, multiple tests will be conducted. When multiple tests are performed without any adjustments the FWER increases. For example, if two independent tests are carried out at the 5% significance level, and the two outcomes are uncorrelated, the probability of finding an intervention effect by chance alone increases to 9.8%. This is increased to 40.1% if ten tests are carried out without adjustment. To maintain the FWER at a pre-specified significance level, adjustments can be made to the p-values (or to the statistical significance level) or a different analysis may be used so that adjustments are not required. Under both approaches, it is necessary to consider the correlations among the outcomes. Selecting a method of analysis that ignores the correlations may lead to adjustments that are too conservative.

The FWER needs to be considered in trials involving multiple primary outcomes when 'success of intervention' is defined as showing an effect on at least one outcome. In this scenario the primary outcomes are referred to as multiple primary outcomes (9) and the p-values must be adjusted for multiplicity. Alternatively the research question in a trial may be formulated so that the 'success of the intervention' is defined as showing an effect on all primary outcomes. In this scenario each outcome is tested at the same significance level without any adjustments for multiplicity. These primary outcomes are called co-primary outcomes (10).

The sample size calculation is an important part of designing a clinical trial. An optimal sample size ensures that the trial is efficient, ethical and cost effective(11). The number of primary outcomes and the correlations among them should be considered when calculating the sample size.

The aim of this study is to identify how multiple primary outcomes are reported and handled in the analysis and sample size calculation of randomised controlled trials recently published

in leading journals. We aimed to assess whether appropriate steps have been taken to safeguard the inferences and to describe problems related to the assessment of multiple outcomes.

Overview of statistical methods

Several methods have been developed to take account of multiple primary outcomes and maintain the FWER at an acceptable level, say 5.0%. The methods include composite primary outcomes, which removes the issue of multiplicity, and p-value adjustment, which adjusts for multiplicity.

A composite outcome is constructed by combining multiple outcomes into a single variable, which offers an overall measure of the health of a patient. For example, the primary composite outcome in a trial might be the time until either a nonfatal ischemic stroke, fatal ischemic stroke or early death after randomisation. Composite outcomes take account of multiplicity without the need to adjust p-values as only one test is performed (12). However, the clinical appropriateness of a composite outcome may be questionable when the intervention appears to affect individual outcomes differently (13).

Several methods have been proposed to adjust p-values, or significance levels, to account for multiplicity, including those of Bonferroni, Šidák, Holm, Hochberg and Hommel. A summary of these methods is provided below and further details can be found elsewhere (14)(15).

Bonferroni's adjustment is an approximate method based on the probability of obtaining a false positive when the outcomes are uncorrelated. It is a simple method where the significance level is divided by the number of primary outcomes. That is, if α is the original, unadjusted, significance level and there are m hypotheses (in this scenario m outcomes) then the Bonferroni's adjusted significance level is $\tilde{\alpha} = \frac{\alpha}{m}$. Šidák's adjustment (16) is the exact version of the Bonferroni adjustment that uses the significance level $\tilde{\alpha} = 1 - (1 - \alpha)^{1/m}$.

Holm's adjustment (17) involves a step-down procedure, whereby the p-values are ordered and successively larger p-values are compared to a successively larger significance level. That is, if the unadjusted p-values are ordered from smallest to largest (i.e. $p_1 \leq p_2 \leq \dots \leq p_m$), and the corresponding ordered null hypotheses are labelled $H_{(1)}, \dots, H_{(m)}$, then Holm suggests rejecting $H_{(i)}$ when for all $j = 1, \dots, i$ $p_i \leq \frac{\alpha}{m-j+1}$. Hochberg's adjustment(18) is a step-up procedure in which successively smaller p-values are compared to increasingly rigorous significance levels. (19). Hommel's method(20) is also a step-up procedure which is more powerful than the Hochberg procedure (14).

Alternatively, multivariate methods can be used to allow multiple outcomes to be simultaneously analysed using a single model (21). These methods are likely to increase the efficiency in estimation (22) compared to analysing outcomes separately. For example, multivariate analysis of variance (MANOVA) (23) may be used to determine if there are differences between trial arms with respect to multiple continuous outcomes, and provides a single p-value to test the overall effect. A follow-up analysis may then consider the effect on each outcome separately, though multiplicity adjustments will need to be made.

Methods

High impact factor journals which publish trials on neurological and psychiatric disorders were selected. Randomised trials are common for these disorders and simple outcomes do not satisfactorily describe the impact of treatments. We hand searched six leading journals featuring neurology and psychiatry studies: The New England Journal of Medicine; The Lancet; The American Journal Psychiatry; JAMA Psychiatry; The Lancet Neurology and Neurology, for reports of randomised trials published between July 2011 and June 2014 inclusive. These journals were selected as they are high impact journals that frequently publish randomised trials in psychiatry and neurology. The impact factor was based upon those published in 2010, the last impact factor year prior to the years for which the data was extracted. Additional supplementary material, including protocols and appendices were reviewed, if they were referred to in the paper.

The following trials were excluded from the analyses: proof of principle trials, phase II trials, including pilot trials and small crossover trials, and secondary analyses of trials. A study was classified as a pilot if it was clearly defined as such, or if it was described, within the discussion section, as an exploratory study prior to a larger study.

For each trial we examined the results in the abstract and the methods used for sample size calculation and statistical analysis. We recorded the number of primary and secondary outcomes and the methods used to account for multiple primary outcomes. An outcome was viewed as primary if it was explicitly stated as such or if it was clearly implied in the aims of the trial. Otherwise, we assumed that all presented outcomes were primary. In the event that the primary outcomes differed in the abstract to the main text, we used the outcomes reported in the main text.

The initial assessments were carried out by one assessor (VV). For those trials where the results were not easily determined, the trials were appraised independently by other assessors (GA and RO). All discrepancies were resolved by discussion between assessors. Statistical analyses were performed using Stata version 12 (24).

Results

From the six journals, we reviewed a total of 3277 abstracts and identified 209 randomised controlled trials that met the inclusion criteria. Details of the study screening process can be seen in Figure 1. The majority of the trials (92%) were parallel-design, individually randomised trials, with a median number of subjects of 242 (IQR 112-549) and a median follow up time of 6 months (IQR 3-17.5 months); Table 1 and Figure 2 summarise the characteristics of these trials. A list of included studies can be found in Additional file 1.

Figure 1: Study screening process.

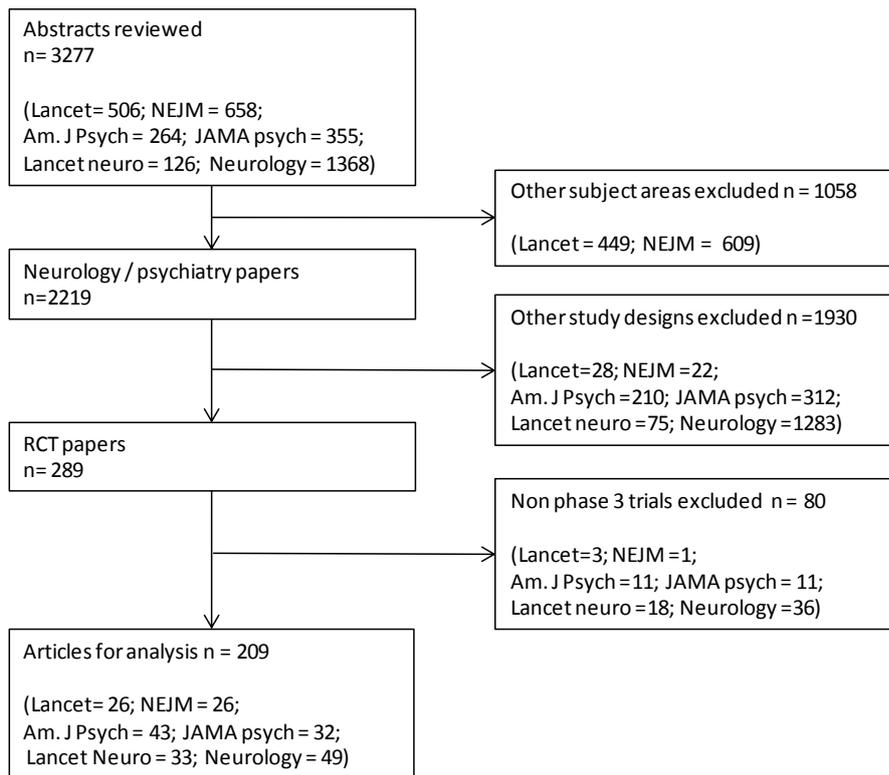
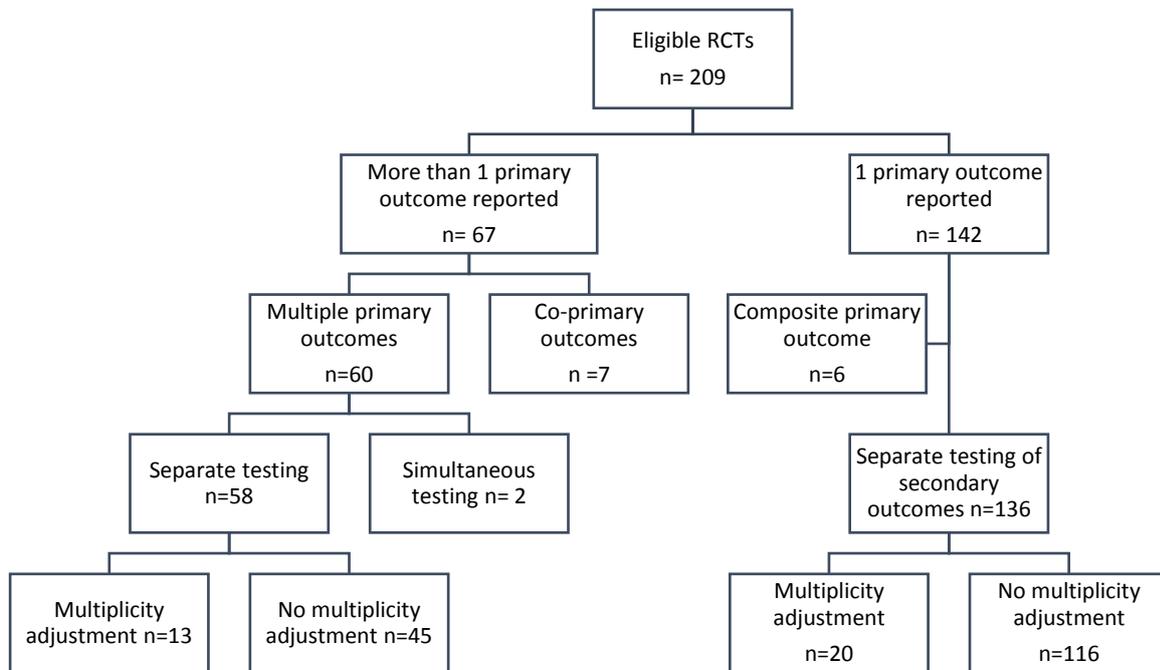


Table 1: Description of the trials included in the study

Results are number (percent).

Characteristic		Number (%) of RCTs
		Total n = 209
Journals	The New England Journal of medicine	26 (12)
	The Lancet	26 (12)
	The American Journal of Psychiatry	43 (21)
	JAMA psychiatry	32 (15)
	The Lancet neurology	33 (16)
	Neurology	49 (24)
Arms per trial	2	144 (68)
	3	52 (25)
	4+	15 (7)
Sites per trial	Single centre trial	36 (17)
	Multi-centre trial	173 (83)
Design of trial	Individually randomised, parallel design	193 (92)
	Individually randomised, factorial	4 (2)

Figure 2: Flow chart of multiplicity within the randomised controlled trials



Trials with no stated primary outcome or with multiple primary outcomes

Of the 209 trials, six (3%) did not clearly specify a primary outcome. These trials did not follow the International Standards for Clinical Trials Registries produced by the World Health Organisation which states that both the primary and secondary outcomes should be defined and pre-specified (25). We therefore, assumed that all outcomes in these trials were equally important and were considered as primary outcomes.

Overall, about a third of the trials (n=60, 29%) reported multiple primary outcomes . Forty-five (75%) of these 60 trials did not include adjustments for multiple comparisons. If multiplicity had been accounted for using Bonferroni’s adjustment, 6 of the 26 trials that reported an effective intervention would have drawn different conclusions.

The other 15 (25%) trials accounted for multiple testing: 6 used Bonferroni’s correction, 7 used other correction methods (Holm, Hochberg-Benjamini, Šidák, Dunnett and sequential adjustments), and 2 performed MANOVA. One justification provided for not accounting for multiple comparisons was “to prevent Type II error”(26).

The abstracts differ by how much the authors have specified the multiple outcomes to be primary.

Just over half (57%, n=34) of the trials clearly specified multiple primary outcomes in the abstract. The remaining abstracts described the outcomes in the abstract without distinction, even though there were later distinguished as primary and secondary in the body of the papers.

Fourteen (23%) of the 60 trials that reported multiple primary outcomes used only one outcome in the sample size calculation. Fourteen of the trials clearly reported sample size calculations that incorporated more than one outcome. Sample size adjustments included a multiplicity-adjusted significance level within the calculation, and calculating the sample size for all the primary outcomes then selecting the largest value (27).

Trials with co-primary outcomes

Seven (3%) of the 209 trials reported co-primary outcomes. Even though it is not necessary, two of these trials accounted for multiple testing of the co-primary outcomes in their analysis: 1 used Hochberg's correction and 1 used a pre-specified testing hierarchy. The abstracts of the seven trials clearly specified all of the co-primary outcomes. When calculating the sample size, 5 trials performed calculations based on all co-primary outcomes and two trials performed calculations based on just one outcome.

Trials with one stated primary outcome

The remaining 142 (67%) trials reported one primary outcome. Of these trials, five reported one primary and no secondary outcomes. Another six (3%) used a composite primary outcome.

Psychiatry, neurology and general medicine journals

When viewing the results by disease area, the trials published in the psychiatric journals most frequently reported more than one primary outcome, with 35 (47%) trials reporting multiple primary outcomes. This compared with 18 trials (22%) in the neurological and 7 (13%) in the general medicine journals. Of those trials analysing multiple primary outcomes, 27 (77%) and 15 (83%) of the trials in the psychiatric and the neurological journals respectively did not account for multiplicity compared to 3 (43%) trials in the general medicine journals.

Drug versus non-drug trials

134 (64%) trials evaluated drug treatments, of which approximately a quarter (30 (22%)) reported multiple primary outcomes and of which only 6 (20%) accounted for multiplicity. A larger proportion (30 (40%)) of the 75 non-drug trials analysed multiple primary outcomes and 9 of these (29%) accounted for multiplicity.

Secondary outcomes

Even though adjustments may be less important for secondary outcomes (8), 19 (13%) of the 142 trials that reported one primary outcome accounted for multiplicity in their secondary outcomes by adjusting the p-values. An additional seven (5%) trials highlighted main secondary outcomes. Five of the 67 (7%) trials that had multiple primary outcomes or co-primary outcomes adjusted for multiple comparisons in the analysis of their secondary outcome: two trials used sequential testing and three trials adjusted the p-values.

Discussion

This review has identified a lack of consistency in the reporting and analysis of multiple outcomes in randomised controlled trials. We often found it difficult to determine from the report the number of primary outcomes being analysed in a trial and there was a lack of consistency when specifying the primary outcome.

A variety of methods to handle multiple primary outcomes were noted in this review. The majority of authors who accounted for multiplicity did so by adjusting the p-values. The most common technique observed was the Bonferroni adjustment. If the outcomes are correlated, the adjustments methods observed in this review are conservative.

In recent years, more complex multivariate methods have been used that utilise the correlations between the outcomes (3). The MANOVA approach was the only multivariate method used despite the fact that multivariate methods could increase the power (3, 28) and reduce the sample size required. Another multivariate method that may be used is the multivariate multilevel model introduced by Goldstein (21). This method uses multivariate multilevel linear modelling which can be implemented when there are multiple outcomes (29-31). This method is an extension of multilevel modelling, where individuals are clustered within larger units, for example hospitals or general practices. This approach views multiple outcomes as repeated measures clustered within individuals and models the correlations between outcomes. The information gained from the correlations may provide more efficient estimates of treatment effect (32), resulting in smaller standard errors. As a result, fewer individuals may need to be recruited when using this method. This method can produce a p-value for each outcome. Following this analysis, multiplicity adjustments would still need to be made. Alternatively, a global treatment effect can be computed that does not need to be adjusted for multiplicity.

In the majority of cases, the trials did not detail any steps to safeguard the inferences made when using multiple primary outcomes. Of these trials, 26 reported significant results. However, six of these would have drawn different conclusions if a Bonferroni adjustment had been applied: one intervention would no longer be effective for any of the primary outcomes and in five trials the intervention would be effective for a small subset of the primary outcomes.

Paradoxically, some of these authors highlighted their awareness of the problems associated with multiplicity by stating they did *not* adjust for multiple outcomes. One reason given for not using any adjustments was “to prevent type II error” (26) whilst others did not provide any justification.

In many of the trials observed, the abstract correctly identified the primary outcomes. However, there is still considerable room for improvement. A number of abstracts provided incomplete descriptions. Some abstracts discussed many outcomes and the corresponding p-values were discussed without any distinction between the primary and secondary outcomes.

This study focuses on neurology and psychiatry trials. Multiple primary outcomes can also be found in other disease areas although we cannot be sure that this situation is the same in these areas. Also, we focused on high quality, high impact journals, so this review could potentially underestimate the proportion of all trials that fail to take account of multiplicity.

Recommendations for reporting

Once the authors have selected the methods for the sample size calculation and analysis, ideally the protocol and journal article should be written in sufficient detail to ensure the reader is fully aware of the methods used. The trial objectives should be clearly stated, as advised in the current ICH guidelines (33). Furthermore, the authors should ensure that they have specified the primary and secondary outcomes, methods of measurements and time points of interest at the start of the trial, as detailed in the International Standards of the World Health Organisation (25). The documentation of these pre-specified outcomes is encouraged by the CONSORT checklist (34). The sample size calculation should be based on the primary outcomes (35).

With regards to multiplicity arising from multiple outcomes, CONSORT state that “authors should exercise special care when evaluating the results of trials with multiple comparisons” (34). We consider that this guidance is limited and recommend that the chosen method to maintain the FWER at the desired level is defined and justification for the choice provided. If the randomised controlled trial is viewed as confirmatory, we agree with the ICH E9 that any aspects of multiplicity “should be identified in the protocol; adjustment should always be considered and the details of any adjusting procedure, or an explanation of why an adjustment is not thought to be necessary, should be set out in the analysis plan” (33). The abstract should be “clear, transparent, and sufficiently detailed” (36), as explained in the CONSORT statement, as readers often base their assessment of the trial on the information provided in the abstract. Consequently, it is important that the abstract is an accurate record of the trial and is not in any way ambiguous or misleading.

Authors should report the sample size calculation and state which outcomes are used in the calculation to ensure the reader is aware of how the study is powered.

There is no general consensus regarding the importance of secondary outcomes; they can be viewed as supportive evidence or as a basis for additional claims. If secondary outcomes are viewed as supportive measurements then although statistical adjustments may not be required “appropriate caution and de-emphasis needs to be exercised in interpreting their results” (37). One option to ensure that secondary outcomes are afforded less emphasis would be to present estimates of the intervention effect for them with the corresponding confidence intervals, rather than the p-values. This would give information about the precision level and if the confidence level included a clinically important effect or not. If the secondary outcomes are used for additional claims then it has been suggested that a procedure is required to deal with them (38). For example, “further confirmatory statistical test on secondary variables can be performed using a further hierarchical order for the secondary variables” (38).

Conclusions

In the selected sample of neurology and psychiatry randomised controlled trials that were published in a number of leading medical journals it was found that multiple primary outcomes were commonly used but often inadequately handled. Unless authors clearly state their chosen primary outcomes and justify their methods of analysis, readers should be wary of unreliable interpretations when multiple primary outcomes are used.

Funding: No funding provided for the specific study.

Contributors: The guarantor (VV) had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Study concept and design: RO, GA, IN, MK, VV. Acquisition of data: VV, RO, GA. Analysis and interpretation of data: VV, RO, GA, MK, IN. Writing of manuscript: VV, RO, GA, MK, IN. Administrative, technical, or material support: VV.

Ethical approval: Not required.

Data sharing: full dataset available from the corresponding author at v.vickerstaff@ucl.ac.uk

References

1. 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):829-38.
2. Blakesley RE, Mazumdar S, Dew MA, Houck PR, Tang G, Reynolds CF, 3rd, et al. Comparisons of methods for multiple hypothesis testing in neuropsychological research. *Neuropsychology*. 2009;23(2):255-64.
3. 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.
4. Tyler KM, Normand SL, Horton NJ. The use and abuse of multiple outcomes in randomized controlled depression trials. *Contemporary clinical trials*. 2011;32(2):299-304.
5. Mayo NE, Scott S. 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*. 2011;40(6):718-24.
6. Pocock S, Geller N, Tsiatis A. The analysis of multiple endpoints in clinical trials. *Biometrics*. 1987;43(3):487-98.
7. Alosch M, Bretz F, Huque M. Advanced multiplicity adjustment methods in clinical trials. *Statistics in Medicine*. 2014;33(4):693-713.
8. Moyé LA. *Multiple analyses in clinical trials: fundamentals for investigators*: Springer Science & Business Media; 2003.
9. Bagiella E. Clinical trials in rehabilitation: single or multiple outcomes? *Archives of physical medicine and rehabilitation*. 2009;90(11):S17-S21.
10. Offen W, Chuang-Stein C, Dmitrienko A, Littman G, Maca J, Meyerson L, et al. Multiple co-primary endpoints: medical and statistical solutions: a report from the multiple endpoints expert team of the Pharmaceutical Research and Manufacturers of America. *Drug Information Journal*. 2007;41(1):31-46.
11. Wittchen H-U, Jacobi F, Rehm J, Gustavsson A, Svensson M, Jönsson B, et al. The size and burden of mental disorders and other disorders of the brain in Europe 2010. *European Neuropsychopharmacology*. 2011;21(9):655-79.
12. ICH E9 Expert Working Group. Statistical principles for clinical trials: ICH harmonised tripartite guideline. *Statistics in Medicine*. 1999;18(15):1903-42.
13. Pogue J, Devereaux P, Thabane L, Yusuf S. Designing and analyzing clinical trials with composite outcomes: consideration of possible treatment differences between the individual outcomes. *PloS one*. 2012;7(4):e34785.
14. Dmitrienko A, D'Agostino R. Traditional multiplicity adjustment methods in clinical trials. *Statistics in Medicine*. 2013;32(29):5172-218.
15. Bretz F, Hothorn T, Westfall P. *Multiple comparisons using R*: CRC Press; 2010.

16. Abdi H. Bonferroni and Šidák corrections for multiple comparisons." In NJ Salkind (ed.). Encyclopedia of Measurement and Statistics. Thousand Oaks, CA: Sage. 2007.
17. Holm S. A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics. 1979;65-70.
18. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika. 1988;75(4):800-2.
19. Pocock S. Clinical trials with multiple outcomes - a statistical perespsective on their design, analysis and interpration. Controlled Clinical Trials. 1997;18:530-45.
20. Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika. 1988;75(2):383-6.
21. Goldstein H. Multilevel statistical models: Wiley. com; 2011.
22. McCulloch C. Joint modelling of mixed outcome types using latent variables. Statistical methods in medical research. 2008;17(1):53-73.
23. Haase RF, Ellis MV. Multivariate analysis of variance. Journal of Counseling Psychology. 1987;34(4):404.
24. StataCorp StataCorp L. College Station, TX: 2011. Stata statistical software: release.12.
25. WHO WHO. International Standards for Clinical Trial Registries. . http://www.who.int/iris/bitstream/10665/76705/1/9789241504294_eng.pdf?ua=1: 2012 November 2012. Report No.
26. Vitiello B, Elliott GR, Swanson JM, Arnold LE, Hechtman L, Abikoff H, et al. Blood pressure and heart rate over 10 years in the multimodal treatment study of children with ADHD. American Journal of Psychiatry. 2014.
27. Odekerken VJ, van Laar T, Staal MJ, Mosch A, Hoffmann CF, Nijssen PC, et al. Subthalamic nucleus versus globus pallidus bilateral deep brain stimulation for advanced Parkinson's disease (NSTAPS study): a randomised controlled trial. The Lancet Neurology. 2012.
28. Yoon FB, Fitzmaurice GM, Lipsitz SR, Horton NJ, Laird NM, Normand SL. Alternative methods for testing treatment effects on the basis of multiple outcomes: simulation and case study. Statistics in Medicine. 2011;30(16):1917-32.
29. Yang M, Goldstein H, Browne W, Woodhouse G. Multivariate multilevel analyses of examination results. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2002;165(1):137-53.
30. Ma X. Stability of school academic performance across subject areas. Journal of Educational Measurement. 2001;38(1):1-18.
31. Mohan J, Twigg L, Taylor J. Mind The Double Gap Using Multivariate Multilevel Modelling to Investigate Public Perceptions of Crime Trends. British Journal of Criminology. 2011;51(6):1035-53.
32. McCulloch C. Joint modelling of mixed outcome types using latent variables. Stat Methods Med Res. 2008;17(1):53-73.
33. Phillips A, Haudiquet V. ICH E9 guideline 'Statistical principles for clinical trials': a case study. Statistics in Medicine. 2003;22(1):1-11.
34. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMC medicine. 2010;8(1):18.
35. Chan A-W, Tetzlaff JM, Altman DG, Laupacis A, Gøtzsche PC, Krleža-Jerić K, et al. SPIRIT 2013 statement: defining standard protocol items for clinical trials. Annals of internal medicine. 2013;158(3):200-7.

36. Hopewell S, Clarke M, Moher D, Wager E, Middleton P, Altman DG, et al. CONSORT for reporting randomized controlled trials in journal and conference abstracts: explanation and elaboration. *PLoS medicine*. 2008;5(1):e20.
37. Pocock SJ. Clinical trials with multiple outcomes: a statistical perspective on their design, analysis, and interpretation. *Controlled clinical trials*. 1997;18(6):530-45.
38. Products CfPM. Points to consider on multiplicity issues in clinical trials. London: The European Agency for the Evaluation of Medicinal Products (EMA). 2002.