Contemporary Clinical Trials 45 (2015) 8–12

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

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

Victoria Vickerstaff [a,b,c,*], Gareth Ambler [b], Michael King [a], Irwin Nazareth [c], Rumana Z. Omar [b]

[a] Division of Psychiatry, University College London, 6th Floor, Maple House, 149 Tottenham Court Road, London W1T 7NF, UK
[b] Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK
[c] The Research Department of Primary Care and Population Health, University College London, Rowland Hill Street, London NW3 2PF, UK

ARTICLE INFO

Article history:
Received 13 May 2015
Received in revised form 18 July 2015
Accepted 20 July 2015
Available online 26 July 2015

Keywords:
Multiplicity
Multiple outcomes
Clinical trials
Neurology
Psychiatry

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.

1. 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 heath 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

* Corresponding author at: Marie Curie Palliative Care Research Department, Division of Psychiatry, University College London, 6th Floor Maple House, 149 Tottenham Court Road, London W1T 7NF, UK.
E-mail addresses: v.vickerstaff@ucl.ac.uk (V. Vickerstaff), g.ambler@ucl.ac.uk (G. Ambler), michael.king@ucl.ac.uk (M. King), i.nazareth@ucl.ac.uk (I. Nazareth), r.omar@ucl.ac.uk (R.Z. Omar).

http://dx.doi.org/10.1016/j.cct.2015.07.016
1551-7144/

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.

2. 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 a is the original, unadjusted, significance level and there are m hypotheses (in this scenario m outcomes) then the Bonferroni's adjusted significance level is $\tilde{a} = \frac{a}{m}$ Šidák's adjustment [16] is the exact version of the Bonferroni adjustment that uses the significance level $\tilde{a} = 1 - (1 - a)^{1/m}$.

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. $p_1 \le p_2 \le \ldots \le p_m$), and the corresponding ordered null hypotheses are labelled $H_{(1)}, \ldots, H_{(m)}$, then Holm suggests rejecting $H_{(i)}$ when for all $j = 1, \ldots, i \cdot \cdot p_i \cdot \cdot \le \frac{a}{m-j+1}$. 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.

3. 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].

4. 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 Fig. 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 Fig. 2 summarise the characteristics of these trials. A list of included studies can be found in Appendix A.

4.1. 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 prespecified [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].