

## **Design, Analysis and Reporting of Multi-Arm Trials and Strategies to Address Multiple Testing**

Ayodele Odutayo\*, DPhil<sup>1,2</sup>, Dmitry Gryaznov\*, PhD candidate<sup>3</sup>, Bethan Copsey, statistician<sup>1</sup>, Paul Monk, registrar<sup>4</sup>, Benjamin Speich, PhD<sup>1,3</sup>, Corran Roberts, statistician<sup>1</sup>, Karan Vadher, statistician<sup>1</sup>, Peter Dutton, statistician<sup>1</sup>, Matthias Briel, MD PhD<sup>3,5</sup>, Sally Hopewell\*, associate professor<sup>1</sup> and Douglas G. Altman\*, director<sup>1</sup> and the ASPIRE study group

\*contributed equally

<sup>1</sup>Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, United Kingdom

<sup>2</sup>Applied Health Research Centre, Li Ka Shing Knowledge Institute of St Michael's Hospital, Toronto, Ontario, Canada

<sup>3</sup>Basel Institute for Clinical Epidemiology and Biostatistics, Department of Clinical Research, University of Basel and University Hospital Basel, Basel, Switzerland

<sup>4</sup>Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, United Kingdom

<sup>5</sup>Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada

### **Correspondence to:**

Ayodele Odutayo  
Applied Health Research Centre,  
Li Ka Shing Knowledge Institute of St Michael's Hospital,  
250 Yonge Street, 6th Floor  
Toronto, ON M5B 2L7

Abstract word count: 320

Manuscript word count: 4167

Keywords: Multi-arm trials, multiple testing, type I error, type II error



## **ABSTRACT**

### **Background**

It is unclear how multiple treatment comparisons are managed in the analysis of multi-arm trials, particularly related to reducing type I (false positive) and type II errors (false negative).

### **Methods**

We conducted a cohort study of clinical trial protocols that were approved by research ethics committees in the United Kingdom, Switzerland, Germany and Canada in 2012. We examined the use of multiple testing procedures to control the overall type I error rate. We created a decision tool to determine the need for multiple testing procedures. We compared the result of the decision tool to the analysis plan in the protocol. We also compared the pre-specified analysis plans in trial protocols to their publications.

### **Results**

Sixty-four protocols for multi-arm trials were identified, of which 50 involved multiple testing. Nine of 50 trials (18%) used a single-step multiple testing procedures such as a Bonferroni correction and 17 (38%) used an ordered sequence of primary comparisons to control the overall type I error. Based on our decision tool, 45 of 50 protocols (90%) required use of a multiple testing procedure but only 28 of the 45 (62%) accounted for multiplicity in their analysis or provided a rationale if no multiple testing procedure was used. We identified 32 protocol-publication pairs, of which 8 planned a global comparison test and 20 planned a multiple testing procedure in their trial protocol. However, 4 of these 8 trials (50%) did not use the global comparison test. Likewise, three of the 20 trials (15%) did not perform the multiple testing procedure in the publication.

**Limitations**

The sample size of our study was small and we did not have access to statistical analysis plans for the included trials in our study.

**Conclusions**

Strategies to reduce type I and type II errors are inconsistently employed in multi-arm trials.

Important analytical differences exist between planned analyses in clinical trial protocols and subsequent publications, which may suggest selective reporting of analyses.

## INTRODUCTION

The prototypical randomized controlled trial (RCT) is a two-arm parallel group trial comparing one intervention to another active intervention or placebo. This design is simple and often used to evaluate stand-alone interventions. However, it may not be sufficient for complex clinical questions. Multi-arm trials – defined as parallel group trials involving three or more treatment arms – are a common alternative trial design that allow simultaneous comparison of multiple interventions<sup>1-3</sup>. Estimates vary with respect to the prevalence of multi-arm trials, but a methodological study of RCTs published in core clinical journals in 2009 found that 13% of the trials (221 of 1690) were multi-arm trials.<sup>4</sup>

Despite potential advantages, multi-arm trials are methodologically and statistically more complex than standard two-arm RCTs<sup>3</sup>. This is due to the number of treatment interventions and the potential number of comparisons that investigators can conduct. The issue of multiple treatment comparisons – also known as multiplicity – must be considered in the rationale, design and analysis of multi-arm trials<sup>1-3,5-12</sup>. Importantly, the primary comparisons of interest should be clearly identified and whether and how the overall type I error rate is controlled in the setting of multiple testing should be justified.<sup>3,6,10</sup>

To date, two methodological studies have examined a cohort of published multi-arm trials<sup>4,10</sup>. These authors noted infrequent reporting of pre-specified global comparison tests and adjustment to control type I error<sup>4,10</sup>. However, these studies relied on published articles, and so, may not be wholly representative of the conduct of contemporary RCTs.

Accordingly, we assembled an international cohort of protocols for multi-arm trials approved by research ethics committees (RECs) in the United Kingdom (UK), Switzerland, Germany, and Canada in 2012. We characterized the planned design and analysis of multi-arm

trials, and examined strategies for managing multiplicity. We also assessed whether there was an association between funding source and reporting of the strategy for addressing multiplicity. Finally, we compared the pre-specified analysis plans in protocols to trial publications when possible.

## **METHODS**

### Identification of Included Studies

We included multi-arm trial protocols approved by RECs in the UK (Bristol regional office), Switzerland (Basel, Bern, Geneva, Lausanne, St. Gallen, Thurgau, Ticino, Zurich), Germany (Freiburg), and Canada (Hamilton) in 2012. The participating RECs approved this study or explicitly stated that no ethical approval was required.

Our analysis was limited to multi-arm trials, as previously defined. Two-arm parallel group trials, and crossover trials were excluded. Additionally, we excluded factorial trials because these trials are often designed to examine treatments that are mechanistically independent, thereby obviating the need for multiple testing in most of these trials. Pilot studies were also excluded because they focus on feasibility and recruitment and the study groups are often not formally compared. Therefore, our analysis focused on studies intended to change clinical practice. The process for identifying studies for inclusion is detailed in the supplementary appendix. This study was conducted according to a pre-specified protocol, with relevant amendments documented (Supplementary Appendix 2). This study was not pre-registered.

### Data Extraction

Data extraction for general characteristics of all trial protocols was performed in duplicate. Methodological items specific to UK trials were extracted in duplicate. Due to restrictions regarding access to clinical trial protocols for the Swiss, German and Canadian cohort, methodological items specific to multi-arm trials were extracted from a random sample of 25% of trial protocols in duplicate (Supplementary Methods). We extracted information on the general characteristics of multi-arm trials, including intervention type, number of study centres (single, multiple or unclear), number of treatment arms, planned sample size, and funding source (solely or partially industry funded versus non-industry funded, Supplementary Methods). In addition, the trial structure was determined based on the type of intervention employed in the trial. Broadly, there are three ways in which interventions are incorporated in multi-arm trials: 1) multiple active interventions; 2) combinations of active interventions; 3) different doses of the same intervention. The structure of each multi-arm trial is important to understand the potential number of comparisons and how multiple testing can be managed (Supplementary Methods). Components of the aforementioned trial structures can also be combined, leading to added complexity in multi-arm trials. We constructed a taxonomy for multi-arm trials to further clarify these comparisons (Supplementary Methods).

We also extracted details on the number of planned primary comparisons, whether all study arms were mentioned in the study objectives and the planned primary analysis. With respect to sample size calculation, we assessed whether the planned primary comparison was used in determining the sample size.

Finally, each trial protocol was assessed to determine whether multiple testing would be performed for the primary outcome and, if yes, whether a global test – defined as an omnibus test that simultaneously compares all groups – would be used. We also assessed whether there was any planned use of multiple testing procedures (MTPs, see Supplementary Methods). Briefly, the rationale for the use of an MTP is that each treatment comparison for the primary outcome corresponds to a null-hypothesis. The significance level for testing each null hypothesis is the type I error rate or alpha, conventionally set at 5%. The type I error rate is the probability of incorrectly rejecting the null hypothesis and obtaining a false positive result. In the setting of multiple comparisons, the type I error rate across all comparisons is known as the family wise error rate (FWER): the probability of obtaining at least one false positive result. In a multi-arm trial, each pairwise comparison could be considered as an additional opportunity for a false positive result. Therefore, the analysis of multi-arm trials may need an MTP to keep the FWER at 5% across all primary comparisons of interest. Strategies to control the FWER at 5% include single-step methods that lower the type 1 error across all comparisons (e.g. Bonferroni correction) or ordering the comparisons of interest, with or without an adjustment to the type I error for each comparison <sup>13</sup>. These methods include the Hochberg procedure, fixed-sequence testing or fall back procedures. A detailed description of strategies to manage multiplicity in multi-arm trials is given in the Supplementary Methods.

If a single-step MTP such as the Bonferroni correction was used, we assessed whether the sample size calculation would be adjusted. Adjusting the sample size in the setting of a single-step MTP is important to ensure the trial is adequately powered for the primary comparison and to reduce type II error, incorrectly failing to reject the null hypothesis (false negative) <sup>11,12</sup>.



## Decision Tool to Determine the Need for Adjustment for Multiple Testing

Using a systematic search of the literature (Supplementary Methods), we identified published articles and guidance documents on MTPs <sup>1-3,5,6,10,13-17</sup> and constructed a decision tool to assess whether multi-arm trials should use MTPs to control the FWER in their analysis (Figure 1). Although there is not universal consensus on when the FWER should be controlled in multi-arm trials, the components of our decision tool are similar to a recent narrative review by an independent group of investigators on this topic <sup>3</sup>. Therefore, our decision tool reflects existing consensus on this topic.

Our decision tool involves three decision points (Figure 1). The rationale underpinning each of these decision points is provided in Supplementary Table 1. Importantly, our decision tool focused on multiplicity issues related to the number of intervention arms. There may be multiplicity issues because of multiple primary outcomes and when interim analyses are planned. However, these latter scenarios are not specific to multi-arm trials and there is guidance available regarding these issues <sup>18</sup>. In the instance of a trial with multiple primary outcomes as well as multiple treatment arms, investigators may benefit from using more complex MTPs that preserve type I error throughout their comparisons <sup>13</sup>.

Trials could conceivably be classified in more than one way based on “Decision Point 2” in our tool. In this instance, the decision points under each potential classification were examined to determine if an MTP was needed. Using the decision tool, we identified which multi-arm trials warranted an MTP in their analysis. We compared this assessment with whether the trial protocol reported a plan to use an MTP.

### Comparison to Trial Publications

PubMed and trial registries were searched iteratively from September 2017 to August 2019 for publications of completed multi-arm trials. We searched PubMed and trial registries until August 2019. The search strategy was individualized for each trial protocol in our study. We used a hierarchical search strategy involving a search for the study title, study acronym (if available), combinations of keywords taken from the study title and trial registration number. Where a publication was not identified, we used a combination of keywords, investigator names, and intervention names. We matched the protocol and the publication based on the trial registration number, the ethics application number (if provided in the published article), the authors, the sample size and the study population. We did not contact study investigators. When a corresponding publication was found, we compared reporting on the use of a global comparison test and on the use of MTPs.

### Data Analysis

We provide descriptive statistics for categorical and continuous variables as appropriate. Poisson regression with robust standard errors was used to estimate risk ratios for the association between industry funding and the reporting of: 1) whether a global test was used or it was explicitly stated that no global test was used; 2) whether an MTP was used when deemed necessary based on the decision tool. Poisson regression was used as opposed to logistic regression because the odds ratio from logistic regression over-estimates the risk ratio given the high event rate for these outcomes. These outcomes were of interest as they have direct implications for the results of multi-arm trials and the potential to influence the risk of type I and type II errors. We did not adjust our analysis for clustering due to the small number of trials. This

analysis is therefore exploratory in nature and the results should be considered hypothesis generating.

## RESULTS

Five hundred and eleven protocols for RCTs were reviewed to identify multi-arm trials (Supplementary Figure 1 and 2). Sixty-four (13%) protocols for multi-arm trials were identified, of which 21 were from the UK, 23 from Switzerland, 14 from Germany and 6 from Canada. Sixty-two of the 64 trials were confirmatory in nature. The general characteristics of included trials from each cohort are detailed in Table 1.

### Methodological Items Related to Multi-Arm Trials

Reporting of methodological items related to multi-arm trials is summarized in Table 2. Multiple testing related to the primary outcome was planned in 50 of 64 trial protocols (Table 2), of which 12 of 50 trials (24 %) intended to perform a global comparison test. With respect to controlling the FWER, 10 of 50 trials (20%) that included multiple testing in their analysis used a single-step MTP such as a Bonferroni correction to account for multiple testing, 16 (32%) used an ordered sequence of primary comparisons and 3 (6%) explicitly stated that an MTP would not be used and provided a reason. Twenty-one of 50 trials (38%) either stated that an MTP would not be used (without a rationale) or made no reference use of an MTP. Finally, in the 10 trial protocols where a single-step MTP was used to account for multiple testing, 7 trials (70%) considered an adjustment in their sample size calculation to maintain statistical power. We

repeated our analysis excluding studies with multiple primary endpoints and these results were unchanged.

#### Comparison Between Trial Protocols and the Decision Tool for Multiple Testing Procedures

Based on the decision tool, 45 of 50 multi-arm trial protocols (90%) required an MTP to control the overall type I error rate in their trial. Among these 45 trials, 28 (62%) used an MTP or provided a rationale for their decision not to use an MTP (Table 3). For the remaining 17 of 45 trials that did not use an MTP when it was required by the decision tool, 12 were trials that compared multiple doses of the same intervention and any arm could be superior to conclude efficacy of the intervention (Example 6 in Figure 1 Caption). Four were trials that compared multiple interventions to a common control and the interventions could not be considered to be independent (Example 4 in Figure 1 Caption). One was a trial examining a combination of active interventions.

Five (10%) of 50 multi-arm trial protocols did not require an MTP based on the decision tool. All of these five trials compared multiple interventions versus a common control and did not require an MTP because the interventions could be considered independent. Only one trial provided a reason for not performing an MTP (Table 3).

Of interest, there were 9 adaptive trials in our sample of multi-arm trials, eight of which required multiple testing based on our decision tool. Three (33%) of the 9 trials planned to use an

MTP, 3 (33%) stated that an MTP would not be used but did not provide a reason and 3 (33%) did not state whether an MTP would be used.

#### Association Between Funding Source and Reporting of Analysis Strategies in Multi-Arm Trials

Forty-one industry funded trials and nine non-industry funded trials involved multiple comparisons for their primary outcome. The use of a global test was reported in 21 of 41 trial protocols for industry funded studies (51%) compared with 3 of 9 trial protocols for non-industry funded studies (33%). There was no evidence of an association between funding source and the use of a global test (Risk Ratio [RR] 1.41, 0.59-3.37). With respect to using an MTP when it was deemed necessary by the decision tool, 25 of 40 industry funded trials (68%) and 3 of 5 non-industry funded trials (60%) reported details about how the FWER was controlled in their analysis (RR 1.04, 0.52-2.08).

#### Comparison Between Trial Protocols and Published Articles with Respect to the Use of a Global Test and Multiple Testing Procedures

Of the 64 multi-arm trials, 42 (65%) were published in peer-reviewed journals and 15 (25%) were unpublished. Among the unpublished trials, 15 (68%) were registered in a clinical trial registry. Eight (53%) of these 15 multi-arm trials were listed as completed, 5 (33%) were listed as terminated and 2 (13%) were listed as still recruiting participants. Of the 8 trials listed as completed, 5 provided results in the registry. All of the terminated trials provided results in the trial registry.

Among the 50 multi-arm trial protocols involving multiple testing, corresponding publications were identified for 32 trial protocols. Eight of 32 trial protocols reported the planned use of a global comparison test but only 3 (38%) ultimately reported this test in the final publication and 1 (13%) provided a reason for not including the global test (Supplementary Table 2). All of the remaining 4 studies that did not use the global test in the publication achieved the primary outcome for at least one of the comparisons of their multi-arm trial. Exact p-values were only given in one study.

Six of 32 trial protocols planned to use a single-step MTP and 14 of 32 trial protocols planned to use an ordered sequence of primary comparisons to control the FWER. Three of these 20 trials (15%) did not use an MTP in the corresponding publication and did not provide a reason (Supplementary Table 3). Of the three studies that did not use an MTP in the publication, one was reported as a two-arm trial, one did not meet its primary endpoint despite not performing the MTP and one achieved its primary endpoint at a p-value of  $<0.0001$  for one of its comparisons but did not report a p-value for the remaining primary comparisons.

## **DISCUSSION**

This study provides an overview of the design and analysis of an international cohort of multi-arm trial protocols. There are three key findings. First, when an adjustment for multiplicity was needed based on our decision tool, only 62% of trial protocols accounted for multiplicity in their analysis strategy or provided a reason for their decision not to use an MTP. Second, among 10 trials that intended to use a single-step method to account for multiple testing, only 7 (70%) also reflected this in their sample size calculation to maintain statistical power. Third, there were

important differences in study methods between the trial protocol and the publication of multi-arm trials, which suggests that selective reporting of analyses in multi-arm trials may have occurred.

To our knowledge, there are only two methodological studies on publications of multi-arm trials. Both studies demonstrated that the use of MTPs was infrequent <sup>4,10</sup>. For instance, in a study of 221 published parallel group multi-arm trials, the authors found that 127 trials (58%) planned a global test comparison in their analysis but only 84 (38%) reported these results <sup>4</sup>. Similarly, only 89 (40%) of multi-arm trials including an adjustment method to control type I error in the sample size calculation or in the statistical analysis <sup>4</sup>. These studies are limited by their reliance on trial publications, which may not reflect the conduct of clinical trials and the complex decision making process regarding adjustment for multiple testing. Therefore, analyses of clinical trial protocols are important for determining how design and analytical issues are addressed in multi-arm trials.

While most trials protocols in our study mentioned all intervention arms in the study objectives and planned analysis, important methodological items related to multiple testing were inconsistently reported. For instance, 45 of 50 trials required use of an MTP based on our decision tool but only 28 of 45 trial protocols (62%) ultimately used an MTP in their analysis or provided a rationale for their decision not to use an MTP. The remaining 17 of 45 (38%) trial protocols did not state how or why they would not control the FWER in the context of multiple comparisons. These latter studies may therefore be more prone to type I error and incorrectly identifying novel interventions as efficacious.

Adaptive trials are an extension of multi-arm trials and involve an added level of complexity. These trials are described as “adaptive” because at regular interim analyses,

recruitment into individual intervention arms can be stopped for lack of efficacy or harm. As adaptive trials often simultaneously examine multiple interventions, endpoints and interim analyses, these studies represent a particularly complex case of multiple testing. Indeed, existing guidance recommends a discussion of the need for multiple testing in adaptive trials <sup>19</sup>. Based on our decision tool, which focuses specifically on the number and nature of the interventions, 8 of the 9 adaptive trials in our analysis required an MTP. However, the studies were evenly split with 3 (33%) adaptive trials planning to use an MTP, 3 (33%) studies stating that an MTP would not be used without providing a reason, and 3 (33%) studies not stating whether an MTP would be used. Although our sample size is small, our findings suggest that the uncertainty about the use of MTPs also extends to adaptive trials.

Notably, only five of 50 multi-arm trials were designed with the intention of examining multiple independent hypotheses simultaneously while maintaining statistical power. Indeed, if a multi-arm trial compares two active but mechanistically independent interventions against a control group, there is no requirement to perform an MTP. For instance, a publicly available example of a trial designed in this manner is the Dressings and securements for the prevention of peripheral intravenous catheter failure in adults (SAVE) trial <sup>20</sup>. In that study, investigators randomized participants to securing their peripheral intravenous catheter to usual care with a polyurethane dressing, usual care and an additional supportive border for the peripheral intravenous, usual care and a tissue adhesive and finally usual care and a securement device. Appropriately, these authors did not perform an MTP. Researchers should more frequently consider multi-arm trial designs as a strategy to improve efficiency and address multiple hypotheses without compromising statistical power.



Our study highlights the complex methodological issues in multi-arm trials and the ongoing uncertainty among investigators in how to address multiple testing in the design and analysis of their trial. By operationalizing the decision making process regarding MTPs, our decision tool is an important source of guidance. Although there is not universal consensus on strategies for managing multiplicity, our tool is a reflection of the existing literature on MTPs <sup>1,3</sup>. Perhaps the most important aspect of our decision tool is that adjustment for multiple testing should not be reflexively performed in all multi-arm parallel group trials. Instead, there should be careful consideration of the research question, the nature of the interventions and the number of primary comparisons. These items should ideally be considered at the design stage of clinical trials, in order to allow for possible adjustment of sample size calculations in the event of a decision to use single-step procedures such as the Bonferroni correction.

Finally, when comparing trial protocols and publications, we noted important differences between the two documents. For instance, 15% of trial protocols, which included an MTP in their analysis, did not perform the testing procedure in the publication and did not provide a reason for this change. Accordingly, our study highlights a possible instance of selective reporting of analyses, wherein important differences exist between the planned and reported statistical analyses in clinical trials. This would likely have implications for the statistical significance of trial results, as well as the interpretation of the efficacy of novel interventions. Pre-specification of statistical analyses of multi-arm trials in a published protocol or perhaps within trial registries would therefore be important for reducing selective reporting <sup>21</sup>.

Our study has several strengths. First, we examined clinical trial protocols from multiple jurisdictions, which strengthens the external generalizability of our findings. Second, we conducted the first review of clinical trial protocols for multi-arm trials and we were able to

examine important issues in the design, analysis and reporting of multi-arm trials that were not examined in prior analyses of published trial reports. However, our study has important limitations. First, despite having reviewed 505 trial protocols, only 64 trial protocols for multi-arm trials were available for our analysis. This resulted in a small sample size for our analyses. Second, most multi-arm trials included in our analysis were industry funded, which contributed to imprecise study estimates in our comparisons based on the funding source. Third, we did not have access to the statistical analysis plans of included trials to clarify whether a reason was provided for changes in the analysis between the protocol and publication. Each of these documents would be required to conclusively identify evidence of selective reporting of statistical analyses. Nonetheless, our comparison of protocols and publications focused on the primary outcomes of the included trials and any analytical changes for a primary outcome should be elaborated in the main publication. Fourth, we did not adjust our regression analysis on industry versus non-industry funding for clustering. This was due to the small number of trials.

Taken together, important gaps exist in the planned analytic strategy for some multi-arm trials, which may have implications for interpretations regarding the efficacy of novel interventions. Guidelines for the preparation of trial protocols for multi-arm trials may be beneficial and allow for appropriate decision making regarding type I error and increasing efficiency in clinical trials.

**KEY MESSAGES**

Multi-arm trials are a useful clinical trial design but are methodologically and statistically complex due to the number of possible treatment comparisons.

Strategies to reduce type I and type II errors are inconsistently employed in multi-arm trials.

Selective reporting of analyses may occur in publications of multi-arm trials.

Few multi-arm trials are designed with the intention of examining multiple independent hypotheses simultaneously while maintaining statistical power, highlighting a potential opportunity to increase efficiency in clinical trials.

## REFERENCES

1. EMA. Points to Consider on Multiplicity Issues in Clinical Trials [Internet]. *ema.europa.eu* 2003 [cited 2016 Nov 23]. Available from: [http://www.ema.europa.eu/docs/en\\_GB/document\\_library/Scientific\\_guideline/2009/09/WC500003640.pdf](http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003640.pdf)
2. EMA. Guideline on multiplicity issues in clinical trials [Internet]. EMA, editor. *ema.europa.eu* 2016 [cited 2018 May 2]. Available from: [http://www.ema.europa.eu/docs/en\\_GB/document\\_library/Scientific\\_guideline/2017/03/WC500224998.pdf](http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2017/03/WC500224998.pdf)
3. Li G, Taljaard M, Van den Heuvel ER, et al. An introduction to multiplicity issues in clinical trials: the what, why, when and how. *International Journal of Epidemiology*. 2017 Apr 1;**46**(2):746–755.
4. Baron G, Perrodeau E, Boutron I, Ravaud P. Reporting of analyses from randomized controlled trials with multiple arms: a systematic review. *BMC Medicine*; 2013 Mar 27;**11**(1):1–1.
5. Bender R, Lange S. Adjusting for multiple testing--when and how? 2001 Apr;**54**(4):343–349.
6. Schulz KF, Grimes DA. Multiplicity in randomised trials I: endpoints and treatments. 2005 May;**365**(9470):1591–1595.
7. Proschan MA, Waclawiw MA. Practical guidelines for multiplicity adjustment in clinical trials. *Control Clin Trials*. 2000 Dec;**21**(6):527–539.
8. Boissel J-P. How to deal with multiple treatment or dose groups in randomized clinical trials? Another approach. *Fundam Clin Pharmacol*. 2007 Apr;**21**(2):155–156.
9. Hothorn LA. How to deal with multiple treatment or dose groups in randomized clinical trials? *Fundam Clin Pharmacol*. 2007 Apr;**21**(2):137–154.
10. Wason JMS, Stecher L, Mander AP. Correcting for multiple-testing in multi-arm trials: is it necessary and is it done? *Trials*. 2014 Sep 17;**15**:364.
11. Wittes J. Sample size calculations for randomized controlled trials. *Epidemiol Rev*. 2002;**24**(1):39–53.
12. Lazzeroni LC, Ray A. The cost of large numbers of hypothesis tests on power, effect size and sample size. *Mol Psychiatry*. 2012 Jan;**17**(1):108–114.
13. Dmitrienko A, D'Agostino R. Traditional multiplicity adjustment methods in clinical trials. 2013 Dec 20;**32**(29):5172–5218.

14. Howard DR, Brown JM, Todd S, Gregory WM. Recommendations on multiple testing adjustment in multi-arm trials with a shared control group. 2016 Sep 19.
15. Rothman KJ. No adjustments are needed for multiple comparisons. 1990 Jan;**1**(1):43–46.
16. Freidlin B, Korn EL, Gray R, Martin A. Multi-arm clinical trials of new agents: some design considerations. *Clin Cancer Res*. 2008 Jul 15;**14**(14):4368–4371.
17. Feise RJ. Do multiple outcome measures require p-value adjustment? 2002 Jun 17;**2**:8.
18. FDA. Multiple Endpoints in Clinical Trials Guidance for Industry [Internet]. *fda.gov* 2017 [cited 2018 May 2]. Available from: <https://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm536750.pdf>
19. Food and Drug Administration. Adaptive Designs for Clinical Trials of Drugs and Biologics Guidance for Industry [Internet]. 2018 [cited 2019 Sep 8]. p. 1–36. Available from: <https://www.fda.gov/media/78495/download>
20. Rickard CM, Marsh N, Webster J, et al. Dressings and securements for the prevention of peripheral intravenous catheter failure in adults (SAVE): a pragmatic, randomised controlled, superiority trial. *Lancet*. 2018 Aug 4;**392**(10145):419–430.
21. Dwan K, Altman DG, Clarke M, et al. Evidence for the selective reporting of analyses and discrepancies in clinical trials: a systematic review of cohort studies of clinical trials. Dwan K, Altman DG, Clarke M, et al., editors. *PLoS Medicine*. 2014 Jun 24;**11**(6):e1001666.

## ACKNOWLEDGEMENTS

We thank the United Kingdom Health Research Authority, who provided access to study protocols.

### ASPIRE STUDY GROUP COLLABORATORS

*Basel Institute for Clinical Epidemiology and Biostatistics,  
Department of Clinical Research, University Hospital Basel*

PD Dr. med. Lars G. Hemkens, MPH

Dr. med. Stefan Schandelmaier, PhD

PD Dr. med. Benjamin Kasenda, PhD

Kimberly Alba Mc Cord, MSc

Prof. Dr. med. Alain Nordmann

*University of Basel*

Dr. med. Alain Amstutz

*Department of Clinical Research*

*University Hospital Basel*

Dr. med. Ramon Saccilotto, MAS

*Clinical Trial Unit, Department of Clinical Research*

*University Hospital Basel*

Belinda von Niederhäusern, PhD

Prof. Dr. med. Christiane Pauli-Magnus

*Cochrane Switzerland*

Dr. med. Erik von Elm, MPH

Dr. med. Elena Ojeda Ruiz, MPH

*Epidemiology, Biostatistics and Prevention Institute (EBPI)*

*University of Zurich*

Yuki Tomonaga, PhD

Prof. Matthias Schwenkglenks

Cochrane Germany

Universitätsklinikum Freiburg:

*Cochrane Germany*

*Universitätsklinikum Freiburg*

Prof. Dr. med. Jörg Meerpohl

Dr. rer. nat. Anette Blümle

Karin Bischoff, MSc

Katharina Kunzweiler, MPH

*Cochrane Germany*

*Universitätsklinikum Freiburg &  
Community Medicine, Dept. Epidemiology of Health Care und Community Health, University  
Medicine Greifswald*  
Laura Rehner, M.A.

*McMaster University*  
Prof. Jason W. Busse, DC, PhD  
Prof. Dominik Mertz, MD, MSc  
Jacqueline Wong, MD, MSc, PhD  
Ngai Chow, DC, MSc (candidate)  
Patrick Ji Ho Hong, MD (candidate)

## **FUNDING**

This study was unfunded.  
BS is supported by an Advanced Postdoc.Mobility grant from the Swiss National Science  
Foundation (P300PB\_177933).

## **CONFLICT OF INTEREST**

None

## **AUTHOR CONTRIBUTIONS**

*Conception and design of the study:* AO, SH, DGA

*Data acquisition and analysis:* All authors

*Interpretation of data for the work:* All authors

*Drafting the manuscript:* AO

*Critical revisions for important intellectual content:* All authors

*Full access to all of the data in the study and takes responsibility for the integrity of the data and  
accuracy of data analysis:* AO

*Final approval of the study and agrees to be accountable for all aspects of the work:* All authors

AO and DG had full access to all the data in the study. AO takes responsibility for the integrity  
of the data and the accuracy of the data analysis

**Table 1: General Characteristics of Multi-Arm Trials**

	<b>All Protocols n (%) (n=64)</b>	<b>United Kingdom n (%) (n=21)</b>	<b>Switzerland n (%) (n=23)</b>	<b>Germany n (%) (n=14)</b>	<b>Canada n (%) (n=6)</b>
<b>Intervention</b>					
Drug	49 (77)	14 (66)	18 (78)	12 (86)	5 (83)
Device	4 (6)	1 (5)	2 (9)	1 (7)	0 (0)
Procedure/Surgery	7 (11)	4 (18)	3 (13)	0 (0)	0 (0)
Behavioural/Lifestyle/ Education	4 (6)	2 (10)	0 (0)	1 (7)	1 (17)
<b>Common Specialties</b>					
Anaesthetics	3 (5)	1 (5)	2 (9)	0 (0)	0 (0)
Cardiology	4 (6)	0 (0)	3 (13)	0 (0)	1 (17)
Endocrinology	4 (6)	3 (14)	0 (0)	1 (7)	0 (0)
Gastroenterology	7 (11)	0 (0)	4 (18)	1 (7)	2 (33)
Oncology	6 (9)	2 (10)	3 (13)	1 (7)	0 (0)
Neurology	9 (14)	4 (19)	2 (9)	3 (21)	0 (0)
Ophthalmology	3 (5)	2 (10)	1 (4)	0 (0)	0 (0)
Orthopedics	2 (3)	2 (10)	0 (0)	0 (0)	0 (0)
Respirology	8 (13)	4 (19)	0 (0)	2 (14)	2 (33)
<b>Study Centres</b>					
Single	5 (8)	1 (5)	3 (13)	1 (7)	0 (0)
Multiple	59 (92)	20 (95)	20 (87)	13 (93)	6 (100)
Unclear	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
<b>Number of Arms</b>					
3	38 (59)	15 (71)	12 (52)	7 (50)	4 (67)
4	14 (22)	2 (10)	5 (22)	5 (36)	2 (13)
>4	12 (19)	4 (19)	6 (26)	2 (14)	0 (0)
<b>Blinding</b>					
Open Label	9 (14)	2 (10)	6 (26)	1 (7)	0 (0)
Blinding Used (Single, Double or Triple Blind)	54 (84)	18 (86)	17 (74)	13 (93)	6 (100)
Unclear	1 (2)	1 (5)	0 (0)	0 (0)	0 (0)
<b>Registered in a Clinical Trial Registry</b>					
Yes	60 (94)	19 (90)	21 (91)	14 (100)	6 (100)
No	4 (6)	2 (10)	2 (9)	0 (0)	0 (0)
<b>Sample size</b>					
Median (Inter-quartile range)	376 (206-630)	420 (200-1000)	300 (156-480)	530 (231- 690)	487 (355- 1023)
<b>Funding</b>					
Non-Industry Funded	11 (17)	5 (29)	5 (22)	1 (7)	0 (0)
Industry Funded	53 (83)	16 (76)	18 (78)	13 (93)	6 (100)



\*Other countries include Germany and Canada. There were only 6 studies from Canada and therefore they were combined with Germany, which had 14 studies.

**Table 2: Methodological Items Related to the Design of Multi-Arm Trials**

	<b>All Protocols n (%) (n=64)</b>
<b>Number of Primary Outcomes</b>	
One	55 (86)
More than one	9 (14)
<b>Allocation Ratio</b>	
Equal	47 (73)
Unequal, reason provided	6 (9)
Unequal, no reason provided	8 (13)
Not stated	3 (5)
<b>Adaptive Trial Design</b>	
Yes	9 (14)
No	55 (86)
<b>Structure</b>	
Combination of Active Treatments	10 (16)
Multiple Independent Unrelated Active Interventions Versus a Common Control	11 (17)
Multiple Doses or Intensities of the Same Intervention	43 (67)
<b>Taxonomic Classification of Multi-Arm Trials</b>	
Combination of Active Treatments	
A:C:AC	3 (5)
A:B:AB	2 (3)
A1:A1B:A2:A2B	2 (3)
A1B:A2Pa:PbB:PaPb***	1 (2)
A1:A2:A3:AB1:AB2:AB3	1 (2)
A:AB:C	1 (2)
Multiple Independent Unrelated Active Interventions Versus a Common Control	
A:B:C	6 (10)
A:B:O	2 (3)
A:B:P or A:B:C:P	2 (3)
AC:BC:ABC:DC:C	1 (2)
Multiple Doses or Intensities of the Same Intervention	
A1:A2:P	21 (32)
A1:A2:A3:P	7 (11)
A1:A2:B	5 (8)
A1:A2:A3:A4:P	4 (6)
A1:A2:B:P	2 (3)
A1:A2:B1:B2A1:A2:P1:P2**	1 (2)
A1:A2:A3:A4:A5:B	1 (2)

A1:P1; A2:P2;A3:P3;A4:A6**	1 (2)
<b>All Arms Mentioned in Study Objectives</b>	
Yes	44 (69)
No	20 (31)
<b>All Arms Mentioned in Planned Primary Analyses</b>	
Yes	50 (78)
No	14 (22)
<b>Planned Primary Comparison Mentioned in Sample Size Calculation</b>	
Yes	58 (91)
No, Planned Primary Comparison Not Mentioned in Sample Size Calculation	6 (9)
No Sample Size Calculation Reported	0 (0)
<b>Interim Analysis Planned</b>	
Yes	29 (45)
No	35 (55)
<b>Multiple Testing Present</b>	<b>(n=64)</b>
Yes	50 (78)
Unclear if Multiple Testing	2 (3)
No, Dose Response Analysis	2 (3)
No, Global Test Only	1 (2)
No, Multiple Arms of Intervention Collapsed to Form 2-arm Trial	8 (13)
No, Only Two Groups Compared and Other Groups are Descriptive	1 (2)
<b>If Multiple Testing, Is a Global Comparison Test Intended?</b>	<b>(n=50)</b>
Yes	12 (24)
Explicitly Stated that No Global Test to be Used	12 (24)
Not Stated in Protocol	26 (52)
<b>If Multiple Testing, Any Planned Strategy to Control the Family Wise Error Rate (FWER)</b>	<b>(n=50)</b>
Yes, Single Step Procedure to Control FWER (e.g. Bonferroni)	10 (20)
Yes, Ordered Sequence of Comparisons	16 (32)
No Strategy to Control FWER Planned AND Reason Provided	3 (6)
No Strategy to Control FWER Planned but Reason <b>NOT</b> Provided	10 (20)
Not Stated How and If FWER Would be Controlled	11 (22)

<b>If Multiple Testing and Type I Error Rate Adjusted, Was the Sample Size Calculation Also Adjusted To Maintain Statistical Power</b>	<b>(n=10)</b>
Sample Size Adjusted or Reason Provided for Decision not to Adjust	7 (70)
No Sample Size Adjustment Used and No Reason Provided	3 (30)

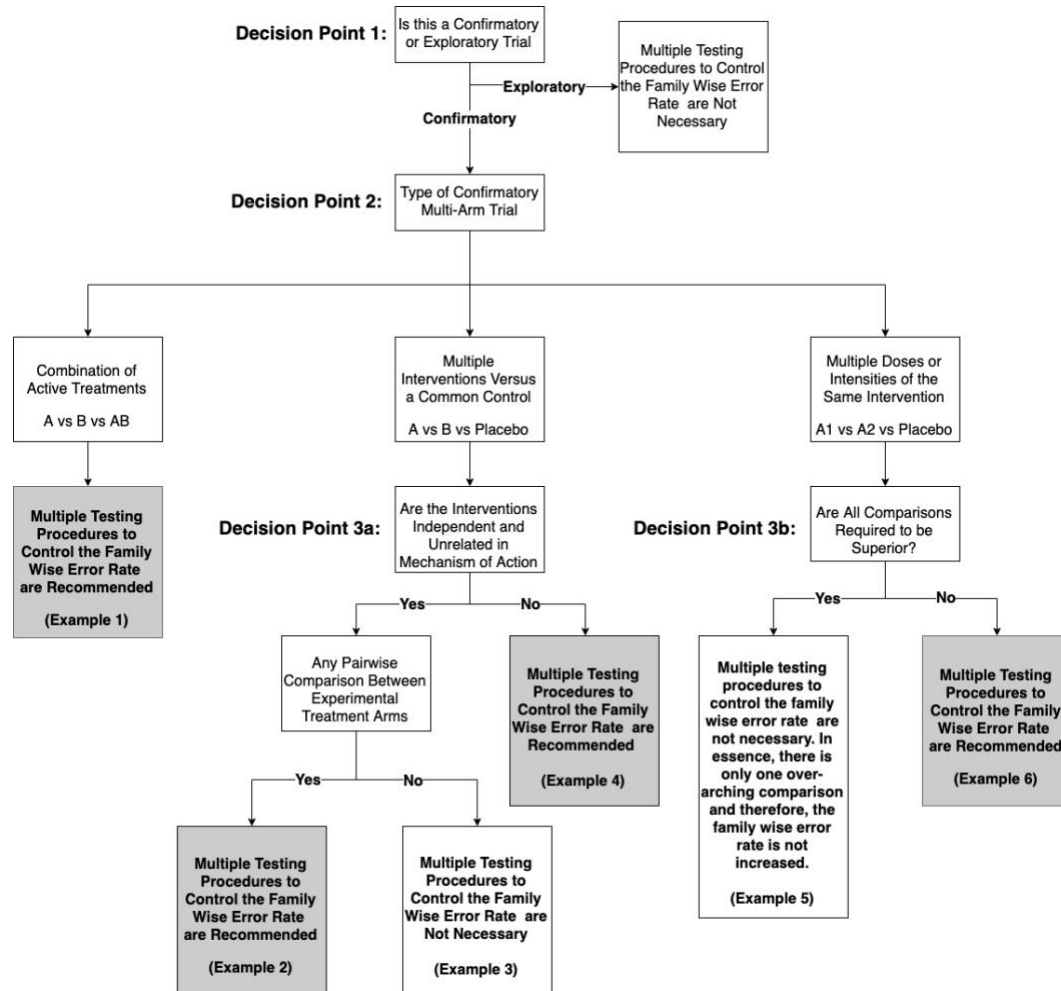
\* where A, B are different treatments, C is control or treatment as usual, A1 and A2 are different doses of the same treatment, AB are a combination of two treatments, P is placebo, O is no treatment and colons separate different treatment arms of the trial. \*\*P1, P2 and P3 represent placebos to correspond to different doses of the same intervention. \*\*\*Pa and Pb correspond to different routes of administration for a placebo.

**Table 3: Comparison Between Analysis Strategy in Trial Protocols and the Decision Tool to Control the Family Wise Error Rate in Multi-Arm Trials with Multiple Testing Planned**

	<b>Yes, Type I Error Adjustment Recommended n (%)</b>	<b>No Type I Error Adjustment Recommended n (%)</b>
<b>All Protocols Combined (n=50 protocols with multiple testing)</b>	<b>(n=45)</b>	<b>(n=5)</b>
Yes, Single Step Procedure to Control FWER* (e.g. Bonferroni)	10 (22)	0 (0)
Yes, Ordered Sequence of Comparisons	16 (36)	0 (0)
No Strategy to Control FWER Planned AND Reason Provided	2 (4)	1 (20)
No Strategy to Control FWER Planned but Reason <i><b>NOT</b></i> Provided	8 (18)	2 (40)
Not Stated How and If FWER Would be Controlled	9 (20)	2 (40)

\*FWER is the family wise error rate

**Figure 1: Decision Tool to Determine the Need for Multiple Testing Procedures in Multi-Arm Trials**



Instances in which adjustment for multiplicity is necessary is indicated in grey boxes. Comparisons of interest are those for the primary outcome. Detailed rationale for the flow diagram is provided in Supplementary Methods.

Example 1: A trial comparing Aspirin, a direct oral anticoagulant and a combination of aspirin and a direct oral anticoagulant for reducing cardiovascular events in adults with atrial fibrillation and stable coronary artery disease.

Example 2: A trial comparing Aspirin, a fish oil supplement and placebo for primary prevention of cardiovascular events in adults. Aspirin and fish oil are compared to placebo as well as one another.

Example 3: A trial comparing Aspirin, a fish oil supplement and placebo for primary prevention of cardiovascular events in adults. Aspirin and fish oil are compared to placebo but not one another.

Example 4: A trial comparing Aspirin, Plavix and placebo for primary prevention of cardiovascular events in adults.

Example 5: A trial comparing one dose of SGLT2 inhibitor, a second dose of SGLT2 inhibitor and placebo for secondary prevention of cardiovascular events in adults. Both doses of SGLT2 inhibitor are expected to be superior to conclude efficacy.

Example 6: A trial comparing one dose of SGLT2 inhibitor, a second dose of SGLT2 inhibitor and placebo for secondary prevention of cardiovascular events in adults. Both dose of SGLT2 inhibitor are not expected to be superior to conclude efficacy.