Multiple Statistical Comparisons in Systematic Reviews: A Quantification of the Issue in Reviews of Anaesthesiological Interventions

Study Protocol

Georgina Imberger[1]
Alexandra Damgaard Vejlby[2]
Sara Bohnstedt Hansen[2]
Jørn Wetterslev[3]

[1]Cochrane Anaesthesia Review Group, Rigshospitalet, Copenhagen, Denmark
[2]Herlev Hospital, Herlev, Denmark
[3]Copenhagen Trial Unit, Rigshospitalet, Copenhagen, Denmark

Abstract

Systematic reviews with meta-analyses often contain high numbers of statistical comparisons. This multiplicity may increase the risk of type I error and may result in spurious conclusions. Such multiplicity has received much attention in the context of single trials and adjustment techniques are often used in this context. But in systematic reviews, few attempts have been made to address the problem. This omission is concerning because systematic reviews are often considered to be the highest quality of evidence. The first aim of this project is to quantify the problem of multiplicity within systematic reviews of anaesthesiological interventions. The second aim is to compare the amount of multiplicity in Cochrane reviews with that in non-Cochrane reviews. The current systematic reviews published by the Cochrane Anaesthesia Review Group will be matched with equivalent non-Cochrane reviews. The number of statistical comparisons made in each review will be measured, and a comparison will be made.

Objectives

1. To quantify the problem of multiplicity within systematic reviews of anaesthesiological interventions.
2. To compare the amount of multiplicity in Cochrane systematic reviews with the amount in non-Cochrane systematic reviews.

Background

Inference is the statisticians attempt to simulate real life and this simulation is in-built with error. Systematic error, or bias, can be limited by good design. Random error, or the result of the 'play of chance', cannot be avoided. When a sample population is used to make an inference about the target population at large, random error is inevitable. Instead of avoiding the risk of random error, its presence must be acknowledged, calculated, and controlled. Probability theories merge with philosophy in this area. Decisions are made as to how much random error is acceptable for inference. Statistical testing of a hypothesis can then be designed to give a defined amount of confidence to the results.

The Hippocratic Oath asks that we 'first do no harm'. Medical researchers stand by a similar philosophy when they design an experiment. They don't want say that drug A is better than drug B, when the truth is that it isn't. This is type 1 error and it represents Hippocrates' 'harm'. To reduce such error, it is usual to set the hypothesis in an experiment at null. That is, it is usual to hypothesize that effect of drug A is no different from drug B. In order to conclude that there is a difference, the null hypothesis must at least be rejected. The risk of this rejection being incorrect, and of making a type 1 error, is often set at five percent. This level of type 1 error has traditionally been considered to be appropriately low.

A P-value may be measured during or at the end of the experiment. Assuming that the null hypothesis is correct, the P-value gives the probability of obtaining a result equal to or more extreme than what was observed in the experiment. When it was initially proposed, the P-value was intended as a tool with which to measure the discrepancy between the null hypothesis and the observed data[1]. It is now commonly used as a single figure measure of statistical significance. A 95% confidence level of the observed difference of the outcome measure between the groups should be given as well and defines the range of results in 95% of situations if the experiment were repeated many times. Confidence intervals provide an impression of the size of an effect. Apart from this difference, the calculation of confidence levels is based on the same frequentist theories as P-values.

When multiple comparisons are made, the risk of type 1 error increases. In medical research, such multiplicity is common and there are many sources: multiple outcomes may be compared, the same outcome may be measured at different time points, there may be multiple intervention groups, more that one descriptive parameter may be used to compare a single treatment effect, there may be analyses made of subgroups, or accumulating data may be compared before the final, so-called fixed, sample size has been reached. From a statistical point of view, minimisation of multiplicity is

ideal. But, from a practical point of view, a trial is a complicated and expensive process. And it seems natural, and ethical, to try to gain as much information as possible. In each study, the additional of statistical evaluations, and the potential to gain more information, has to be balanced against the increasing risk of type 1 error.

Apart from minimising the number of comparisons made, techniques can be used to adjust for effects of multiplicity on type one error risk. The Bonferroni method is probably the most well known [2]. In its simplest form, a p-value is calculated for each comparison and then that p value is multiplied by the total number of comparisons. This approach is simple, but it has the limitation of assuming the independence of each comparison. Such independence is rarely met in a single study and can therefore limit the power of a study to find a true difference[3]. In short, the Bonferroni method is often too conservative. Resampling-based procedures use dependencies and distributional characteristics of the test statistics to calculate adjusted p-values. Gate-keeping procedures may represent solutions as well. These processes are more complicated, but they incorporate of the inter-dependency of multiple comparisons in a single trial, allowing much more preservation of power [4].

The Bonferroni method and resampling-based techniques both provide solutions that are not specific for the source of multiplicity. Other techniques have been developed for particular sources of multiplicity[3]. There are many examples and they include: multivariate techniques which can be used for the comparison of multiple outcome measures[3], multiple stage tests which can be used for the comparison of multiple groups[5] and monitoring boundaries which can be used for interim analyses of accumulating data[3].

A Bayesian approach to analysis may provide a solution for multiplicity. This technique avoids the use of frequentist hypotheses testing, and incorporates an estimation of 'prior odds' of the null hypothesis. The data from the trial is then used to calculate the Bayes factor, or likelihood ratio, which aims to provide a ranking of evidence rooted exclusively in the data of a specific trial. The Bayes factor is then multiplied by the prior odds, giving 'post-study' odds of the null hypothesis. In this process, the probability of misleading evidence can be bounded, independent of the number of comparisons[6] , making the technique less susceptible to the effects of multiplicity.

One of the main concerns with using Bayesian statistics has been in the estimation of the prior odds and the amount of bias that is introduced during such estimation. Supporters of Bayesian techniques argue, however, that a prior estimate of an odds ratio allows the incorporation of previous evidence, and such incorporation is a fundamental part of the gaining of knowledge[7]. Another purported advantage of the Bayesian approach is that it requires that an alternative hypothesis is defined or that the probability of retrieving observed data under all possible alternative hypotheses is calculated. This requirement is in contrast with frequentist techniques, which treat the rejection of the null hypothesis (independent of the alternative) as the goal of an evaluation of the evidence. The importance of this difference is that a Bayesian approach may be able provide a more practical and reliable clinical conclusion.

Systematic reviews present a special case of multiplicity. A systematic review is a study that aims to collate all the reliable evidence available in order to address a specific research question. Meta-analysis refers to the process of statistically combining results from different studies and this process can lead to much multiplicity. First, all the sources of multiplicity in all of the included trials may be relevant. Secondly, the very premise of meta-analysis is re-analysis of accumulating data, thereby multiplying the number of comparisons made. And thirdly, authors of systematic

reviews often aim to cover a topic thoroughly, sometimes with many outcomes, many subgroups and many sensitivity analyses.

Multiplicity has received much attention in the context of single trials and the adjustment techniques mentioned above were all developed in this context. But in systematic reviews, few attempts have been made to address the problem. In the reviews themselves, the presence of multiplicity is rarely mentioned [8]. A recent review on the topic of multiplicity in systematic reviews concluded that the issue requires recognition and further research is required[7].
The Cochrane Collaboration is an international organization that prepares, maintains and promotes systematic reviews. From within this organisation, the issue of multiplicity in systematic reviews has begun to receive attention[9 10]. But this attention has so far been limited and the omission is concerning because systematic reviews are considered to be the highest quality of evidence.

Before solutions are considered and implemented, the size of this issue of multiplicity in systematic reviews deserves clarification. This project aims to assist in that clarification. The presence of multiplicity will be quantified in a population of Cochrane anaesthesiological reviews and in a population of non-Cochrane anaesthesiological reviews. The main aim is to estimate the overall quantity of multiplicity within systematic reviews. A secondary aim is to compare the quantity of multiplicity in Cochrane reviews with that in non-Cochrane reviews.

Whether the issue of multiplicity is addressed in the content of the review will be recorded. The reviews will then be measured for the number of sources of multiplicity and the number of statistical comparisons made. The issue of repeated looks at accumulating data, both from multiple meta-analyses on the same topic and from updating of reviews, will not be addressed in this study.

Methods

Selection of Reviews

The Cochrane Anaesthesia Review Group (CARG) currently has 59 published systematic reviews (November 2009). These are listed in Appendix 1. The CARG reviews which contain meta-analyses will be selected from these 59. For each selected CARG review, a systematic review from a paper journal will be selected.

The reviews from paper journals will be matched to the CARG reviews. That is, for each Cochrane review, a systematic search will be conducted to look for a non-Cochrane equivalent.

Search strategies will be designed using search terms defining the intervention and the population studied in each Cochrane review. Search strategies will be recorded. In order to achieve sensitivity in our searches, we will use strategies described in the Cochrane Handbook of Intervention. (Higgins 2008), we will use the multiple search engines (Medline, Embase, Central, CINAHL, Web of Science, IndMED and KoreaMED) and we will not apply any language restrictions.

We will attempt to match each Cochrane review with a non-Cochrane review looking at the same intervention. If more than one non-Cochrane review is found with the same intervention, the matching will be based on the participants. If more than one non-Cochrane review is found with the same intervention and the same participants, then the matching will be based on the outcomes. If more than one non-cochrane review found matching participants, intervention and outcomes, then the one published closest to the Cochrane review will be selected.

If we are unable to find a match based on the intervention, we will attempt to do so using the participants. If more than one non-Cochrane review is found matching the participants, then the one published closest to the Cochrane review will be selected.

If a non-Cochrane review cannot be matched to a Cochrane review (ie none can be found with the same intervention or the same study population), the first anaesthesiological review published in the same year as the Cochrane review will be selected. For this process, both Medline and Embase search engines will be used, alternating between each selection.

Non-cochrane reviews with any of the same authors as the matching Cochrane review will be excluded.

The current population of CARG reviews only include randomised controlled trials in their meta-analyses. Therefore, non-Cochrane reviews which include observational studies will be also excluded.

Measurements

Each included review will be assessed for the quantity of multiplicity within that review. The following parameters will be recorded:

(i) The sources of multiplicity present in the paper: multiple outcomes (including multiple time-points and multiple effect measures), multiple groups, subgroup analyses, sensitivity analyses.
(ii) The number of statistical comparisons made.
(iii) Whether a primary outcome is quoted.
(iv) How many primary outcomes are quoted.
(v) Whether the issue of multiplicity is referred to in the study.
(vi) Whether any technique is used to adjust for the issue of multiplicity.
(vii) Whether it is clear from the text of the paper exactly how many comparisons were made.

We aim the measure the number comparisons actually performed in these reviews. Therefore, the measurements will be taken from the results sections in the papers. The number of analyses planned and described in the methods sections will not be measured.

Where it is not otherwise stated, it will be assumed that sensitivity analyses were applied to all outcomes, all multiple group comparisons and all subgroup analyses. That is, when the text does not clearly defined how many comparisons were made, the maximum number, given the description available, will be recorded. We will attempt to clarify the number comparisons based on the

information provided in the published paper. This attempt will include reading any supplementary information referred to in the text. We will not contact any authors of any reviews.

The assessment of a Cochrane review will be alternated with the assessment of a non-Cochrane review.

The measurement process will be conducted by two independent investigators. When there is a discrepancy in measurements, a third investigator will review the results, and will aim to find an agreement between the two investigators. In the case that there are two different and equally valid interpretations of the text, an average of the measurements will be taken.

A copy of the data extraction form is in Appendix 2.

A pilot study will be conducted using the first ten reviews (5 Cochrane reviews and 5 non-Cochrane reviews). This protocol will be reviewed after this pilot phase, and adjusted according to any practical difficulties with data extraction, with a view to optimising the accuracy of the results. The results from these first ten reviews will be included in the final analysis.

Statistical analyses

Primary Outcome

The primary outcome for comparison will be the number of statistical comparisons made (prior to sensitivity analyses).

The number of statistical comparisons made will be compared statistically between the two groups (Cochrane and non-Cochrane). The null hypothesis is that there is no difference in the number of statistical comparisons in Cochrane systematic review than in non-Cochrane systematic reviews. The distribution of numbers of statistical comparisons is unknown. Therefore, a two tailed Man-Whitney test will be used to test the null hypothesis.

Secondary Outcomes

1. The number of statistical comparisons made (including sensitivity analyses)
2. The proportion of studies in which a primary outcome is quoted
3. The proportion of studies in which it is clear exactly how many statistical comparison have been made.
4. The proportion of studies in which the issue of multiplicity is addressed

As for the primary outcome, the number of statistical comparisons made (including sensitivity analyses) will be compared using a two-tailed Man-Whitney test.

The proportions will be compared statistically between the two groups (Cochrane and non-Cochrane). The null hypothesis is that there is no difference. A Fisher exact test will be used to test this null hypothesis.

Sensitivity analyses

We plan to conduct three sensitivity analyses for the primary outcome and for the first secondary outcome:

1. Using only the reviews that were successfully matched with the same intervention.
2. Using only the reviews that were matched with a review that was published within the same 2 years.
3. Excluding the paired reviews with 'complex' interventions. That is, pairs where the Cochrane review involves large numbers of intervention groups. These reviews are inherently more difficult to accurately match.

With regard to the power for the primary outcome, calculations have been made assuming normal distribution of numbers of statistical comparisons made and the use of a two-tailed students' t test. A power calculation yields 90% power to detect a mean difference of 1 test, given a standard deviation of 1.5 tests or 80% power to detect a mean difference of 2 tests if the standard deviation is 3.5 tests.

The issue of multiplicity may be important within our own study. We may choose to update our comparison at a later date, perhaps also adding further outcome measures. We will therefore calculate a Bayes factor for our primary outcomes and this likelihood measurement will be used to estimate post-study odds. We have no pre-study evidence as to whether the number of comparisons is different in Cochrane and non-Cochrane reviews. We will therefore estimate our pre-study odds as 1.

References

1. Fisher R, Statistical Methods and Scientific Inference. 3rd ed. New York: Macmillan; 1973.
2. Bland JM, Altman DG. Multiple significance tests: the Bonferroni Method. *British Medical Journal* 1995; 310: 170
3. Bender R, Lange S. Adjusting for multiple testing – when and how? *Journal of Clinical Epidemiology* 2001; 54: 343-349
4. Westfall PH, Young SS. *Resampling-based multiple testing.* New York: Wiley, 1993
5. Godfrey K. Comparing means of several groups. *New England Journal of Medicine* 1985; 313: 1450-1456.
6. Goodman SN. Introduction to Bayesian methods 1: measuring the strength of evidence. *Clinical Trials* 2005; 2: 282-290.
7. Goodman SN. Toward Evidence-Based Medical Statistics. 1: The P-Value Fallacy. *Annals of Internal Medicine* 1999; 130: 995-1004
8. Biester K, Lange S. The multiplicity problem in systematic reviews [Abstract]. *XIII Cochrane Colloquium, Melbourne, 22-26 October 2005.* Program and Abstracts, 2005.
9. Bender R, Bunce C, Clarke M, Gates S, Lange S, Pace NL, Thorlund K. Attention should be given to multiplicity issues in systematic reviews. *Journal of Clinical Epidemiology* 2008; 61: 857-865.
10. Thorlund K, Devereaux PJ, Wetterslev J, Gyatt G, Ioannidis JPA, Thabane L, Gluud L, Als-Nielsen B, Gluud C. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? *Internation Journal of Epidemiology* 2009; 38(1): 276-286
11. Higgins JPT, Green S, Scholten RJPM. *Cochrane Handbook of Systematic Reviews of Interventions.* Chichester (UK): John Wiley & Sons, 2008.

Appendix 1

CARG Reviews October 2009

1. **Adrenaline (epinephrine) for the treatment of anaphylaxis with and without shock**
2. **Alpha-2 adrenergic agonists for the prevention of cardiac complications among patients undergoing surgery**
3. **Antibiotic prophylaxis for surgical introduction of intracranial ventricular shunts**
4. **Antifibrinolytic agents for reducing blood loss in scoliosis surgery in children**
5. **Antifungal agents for preventing fungal infections in non-neutropenic critically ill patients**
6. **Antithrombin III for critically ill patients**
7. **Beta lactam antibiotic monotherapy versus beta lactam-aminoglycoside antibiotic combination therapy for sepsis**
8. **Bispectral index for improving anaesthetic delivery and postoperative recovery**
9. **Caudal epidural block versus other methods of postoperative pain relief for circumcision in boys**
10. **Central venous access sites for the prevention of venous thrombosis, stenosis and infection in patients requiring long-term intravenous therapy**
11. **Closed tracheal suction systems versus open tracheal suction systems for mechanically ventilated adult patients**
12. **Corticosteroids for treating severe sepsis and septic shock**
13. **Drugs for preventing postoperative nausea and vomiting**
14. **Early extubation for adult cardiac surgical patients**
15. **Epidural pain relief versus systemic opioid-based pain relief for abdominal aortic surgery**
16. **H1-antihistamines for the treatment of anaphylaxis with and without shock**
17. **High-frequency ventilation versus conventional ventilation for treatment of acute lung injury and acute respiratory distress syndrome**
18. **Human recombinant activated protein C for severe sepsis**
19. **Hypothermia for neuroprotection in adults after cardiopulmonary resuscitation**
20. **Incentive spirometry for prevention of postoperative pulmonary complications in upper abdominal surgery**
21. **Inhaled nitric oxide for acute hypoxemic respiratory failure in children and adults**

22. **Inhaled nitric oxide for the postoperative management of pulmonary hypertension in infants and children with congenital heart disease**
23. **Interventions for protecting renal function in the perioperative period**
24. **Intravenous immunoglobulin for treating sepsis and septic shock**
25. **Intravenous versus inhalation anaesthesia for one-lung ventilation**
26. **Lidocaine for preventing postoperative sore throat**
27. **Lung protective ventilation strategy for the acute respiratory distress syndrome**
28. **Non-pharmacological interventions for assisting the induction of anaesthesia in children**
29. **Non-steroidal anti-inflammatory drugs and perioperative bleeding in paediatric tonsillectomy**
30. **Noninvasive positive pressure ventilation as a weaning strategy for intubated adults with respiratory failure**
31. **Nutritional support for critically ill children**
32. **Optimal timing for intravenous administration set replacement**
33. **Paracervical local anaesthesia for cervical dilatation and uterine intervention**
34. **Partial liquid ventilation for preventing death and morbidity in adults with acute lung injury and acute respiratory distress syndrome**
35. **Partial liquid ventilation for the prevention of mortality and morbidity in paediatric acute lung injury and acute respiratory distress syndrome**
36. **Patient controlled intravenous opioid analgesia versus continuous epidural analgesia for pain after intra-abdominal surgery**
37. **Peribulbar versus retrobulbar anaesthesia for cataract surgery**
38. **Perioperative fluid volume optimization following proximal femoral fracture**
39. **Pharmacologic therapies for adults with acute lung injury and acute respiratory distress syndrome**
40. **Premedication for anxiety in adult day surgery**
41. **Pulmonary artery catheters for adult patients in intensive care**
42. **Pulse oximetry for perioperative monitoring**
43. **Recompression and adjunctive therapy for decompression illness**
44. **Recruitment manoeuvres for adults with acute lung injury receiving mechanical ventilation**
45. **Rocuronium versus succinylcholine for rapid sequence induction intubation**
46. **Sedation versus general anaesthesia for provision of dental treatment in under 18 year olds**

47. **Selenium supplementation for critically ill adults**
48. **Simple aspiration versus intercostal tube drainage for primary spontaneous pneumothorax in adults**
49. **Single, double or multiple injection techniques for axillary brachial plexus block for hand, wrist or forearm surgery**
50. **Stimulation of the wrist acupuncture point P6 for preventing postoperative nausea and vomiting**
51. **Sub-Tenon's anaesthesia versus topical anaesthesia for cataract surgery**
52. **Sugammadex, a selective reversal medication for preventing postoperative residual neuromuscular blockade**
53. **Supplemental perioperative steroids for surgical patients with adrenal insufficiency**
54. **Target-controlled infusion versus manually-controlled infusion of propofol for general anaesthesia or sedation in adults**
55. **Topical anaesthesia alone versus topical anaesthesia with intracameral lidocaine for phacoemulsification**
56. **Transient neurologic symptoms (TNS) following spinal anaesthesia with lidocaine versus other local anaesthetics**
57. **Ultrasound guidance for peripheral nerve blockade**
58. **Vasopressors for shock**

Appendix 2

Data Extraction Form

Paper _____

Outcome Measured (Including multiple time points and multiple effect measures)	**Multiple Groups?**	**Subgroup Analyses?**	**Number of comparisons prior to sensitivity analysis**	**Sensitivity Analysis?**	**Total Number of comparisons for this outcome.**

Total number of outcomes	Total number of comparisons before sensitivity analyses	Total number of comparisons

Is a primary outcome quoted? _____

If yes, how many primary outcomes are quoted? _____

Is the issue of multiplicity referred to in the study? _____

Are there any techniques used to adjust for the issue of multiplicity? _____

Is it clear how many statistical comparisons were made in this study? _____