ELSEVIER

Journal of Clinical Epidemiology 61 (2008) 857–865

Journal of Clinical Epidemiology

REVIEW ARTICLE

Attention should be given to multiplicity issues in systematic reviews

Ralf Bender[a,b,*], Catey Bunce[c], Mike Clarke[d], Simon Gates[e], Stefan Lange[a], Nathan L. Pace[f], Kristian Thorlund[g]

[a]*Institute for Quality and Efficiency in Health Care, Dillenburger Street 27, Cologne D-51105, Germany*
[b]*Faculty of Medicine, University of Cologne, Joseph-Stelzmann-Street 20, Cologne D-50931, Germany*
[c]*Research & Development, Moorfields Eye Hospital, City Road, London EC1V 2PD, UK*
[d]*UK Cochrane Centre, Summertown Pavilion, Middle Way, Oxford OX2 7LG, UK*
[e]*Warwick Clinical Trials Unit, Warwick Medical School, Coventry CV4 7AL, UK*
[f]*Department of Anesthesiology, University of Utah, 30 North 1900 East, Salt Lake City, Utah 84132, USA*
[g]*Copenhagen Trial Unit, Copenhagen University Hospital, Blegdamsvej 9, Copenhagen DK-2100, Denmark*

Accepted 7 March 2008

Abstract

Objective: The objective of this paper is to describe the problem of multiple comparisons in systematic reviews and to provide some guidelines on how to deal with it in practice.

Study Design and Setting: We describe common reasons for multiplicity in systematic reviews, and present some examples. We provide guidance on how to deal with multiplicity when it is unavoidable.

Results: We identified six common reasons for multiplicity in systematic reviews: multiple outcomes, multiple groups, multiple time points, multiple effect measures, subgroup analyses, and multiple looks at accumulating data. The existing methods to deal with multiplicity in single trials can not always be applied in systematic reviews.

Conclusion: There is no simple and completely satisfactory solution to the problem of multiple comparisons in systematic reviews. More research is required to develop multiple comparison procedures for use in systematic reviews. Authors and consumers of systematic reviews should give serious attention to multiplicity in systematic reviews when presenting, interpreting and using the results of these reports.

Keywords: Cumulative meta-analysis; Meta-analysis; Multiple comparisons; Simultaneous statistical inference; Statistical data interpretation; Systematic reviews

1. Introduction

Many trials in health care research generate a multiplicity of data, hypotheses, and analyses leading to the performance of multiple statistical tests or the calculation of multiple confidence intervals (CIs). The problem of multiple comparisons is well known in clinical trials [1] and epidemiology [2]. Comparisons made in systematic reviews resemble those made in clinical and epidemiological studies and thus will often generate multiplicity [3]. There is an extensive statistical literature about multiplicity and many statistical approaches have been developed to adjust for multiple comparisons in various situations [4–7]. These methods, however, have received little emphasis in systematic reviews. This paper describes the problem of multiple comparisons in the context of systematic reviews and meta-analyses, and provides some guidance on dealing with it.

2. The problem of multiple comparisons

Authors of systematic reviews may perform multiple comparisons for many reasons, for example, multiple outcomes, multiple intervention groups, or multiple time points. Using the conventional significance level of $\alpha = 5\%$, it is expected that one in 20 independent significance tests will be ''statistically significant'' even when there is truly no difference between the interventions being compared. Likewise, it is expected that one in 20 independent 95% CIs will not include the true parameter value. The more analyses that are done, the more likely it is that some tests will be found to be statistically significant although there is no true effect, or that some CIs will not include

* Corresponding author. Department of Medical Biometry, Institute for Quality and Efficiency in Health Care, Dillenburger Street 27, D–51105 Cologne, Germany. Tel.: +49-221-35685-451; fax: +49-221-35685-891.
E-mail address: ralf.bender@iqwig.de (R. Bender).

0895-4356/08/$ – see front matter
doi: 10.1016/j.jclinepi.2008.03.004

the true parameter value. For instance, after 14 independent significance tests, the probability is more than 50% that at least one test will be significant, even when there is no true effect. This probability is called the experimental error rate (EER) and is calculated by the formula $EER = 1-(1-\alpha)^k$, where k denotes the number of tests [4]. Note that these calculations only hold if the k tests are independent. If the k tests are correlated, no simple formula for the EER exists because EER will depend on the kind and strength of this correlation.

Frequently, the global null hypothesis that all individual null hypotheses are true simultaneously is of limited interest to the researcher. Therefore, procedures for simultaneous statistical inference have been developed which control the maximum experimental error rate (MEER) under any complete or partial null hypothesis [4–6]. Procedures to calculate simultaneous CIs are also available [7]. The MEER is the probability of falsely rejecting at least one true individual null hypothesis, irrespective of which and how many of the other individual null hypotheses are true. A multiple test procedure that controls the MEER also controls the EER but not vice versa. Thus, controlling the MEER is the best protection against false conclusions and leads to the strongest statistical inference.

Whether the application of a multiple comparison procedure is required depends primarily on whether the multiple tests belong to the same, single hypothesis or whether they relate to different questions. If the latter, adjustments for multiple testing may be unnecessary, because it is sufficient to control the significance level for each individual question separately [5]. This decision has to be made in each specific situation and usually there will be different opinions about what constitutes the ''correct'' decision. In general, however, the MEER should be under control when the results of a well-defined family of multiple tests are summarized in one conclusion for the whole hypothesis. Otherwise the type 1 error of the final conclusion is not under control, which would mean that the aim of significance testing is not achieved.

In any research where multiple comparisons are performed, adequate planning of the statistical testing of hypotheses and the adjustments for multiple testing or multiple interval estimation should ideally be done at the design stage. Unfortunately, in systematic reviews the data that will be available for the included studies are usually not known at the outset. This makes a priori planning of multiple comparison procedures more difficult for systematic reviews, and the procedures for simultaneous statistical inference in single studies cannot be transferred automatically for dealing with multiplicity in systematic reviews. The simple Bonferroni procedure [8], which would be feasible also in systematic reviews, rarely represents an adequate solution to the problem of multiple comparisons. In the following, the main areas in which multiplicity problems arise in systematic reviews are summarized and explained.

3. Causes of multiplicity problems in systematic reviews

3.1. Multiple outcomes

Clinical trials often have more than one primary outcome [1]. Response to treatment may be measured in a variety of ways. For instance, a trial studying cancer chemotherapy may study overall mortality, but the quality of life experienced by patients whilst on treatment might also be of importance. Systematic reviews collating results from multiple clinical trials share this feature. Critical review of trial reports can be extremely time consuming and it is therefore common for authors of reviews to express a need to gather as much information as possible. Collating large amounts of data is not a problem (other than by increasing the time and resources needed for the review) but analyzing many outcomes is problematic when the conclusion that an intervention is beneficial is made if *any* outcome reaches statistical significance. For example, Table 1 presents outcomes listed for two protocols published in The Cochrane Library [9, 10]. These lists might expand further during the review process as the authors become aware of other ways that trialists have evaluated the effects of the treatments being investigated. Clearly the authors of these reviews will need some way of dealing with multiplicity if they are to draw reliable conclusions from multiple meta-analyses.

One strategy for reducing problems due to multiple testing in these reviews is to specify a priori in the protocol that whilst all results will be reported, conclusions will be drawn only from meta-analyses regarding a smaller, prespecified subset of outcomes. In the first example [9], this could be the one primary outcome. Alternatively, it might be possible to combine outcomes into meaningful categories such as ''any failure'' and restrict analyses to this combined outcome rather than its component parts (e.g., failure to control intraocular pressure, failure to stabilize visual field progress, etc., in example 2 [10]).

One promising approach to deal with multiple outcomes in systematic reviews is given by multivariate meta-analysis, which has the advantage of providing a complete and concise description of all data simultaneously rather than a number of univariate meta-analyses separately for each outcome [11, 12]. One problem of this approach is that neither the within-study nor the between-study correlations are generally known and future research needs to address this issue [13, 14].

3.2. Multiple groups

Systematic reviews synthesize the effects of an intervention (often called ''the treatment group'') compared to another intervention or control group in a defined set of patients. Sometimes the intervention under investigation is a superset of different treatments, all of which have the same therapeutic goal and are compared against similarly

Table 1
Outcomes listed for two protocols published in the Cochrane Library [9, 10]

Title	Interventions for Mooren's ulcer	Laser trabeculoplasty for open angle glaucoma
Objective	To assess the effectiveness of the various intervention(s) for Mooren's ulcer	To investigate the effects of laser trabeculoplasty for treating open angle glaucoma when compared to medication, incisional glaucoma surgery or no intervention. We will also compare different technologies of laser trabeculoplasty for treating open angle glaucoma
Primary outcome(s)	Complete healing—present or not, after 8 weeks of follow up	1) Failure to control intraocular pressure 2) Failure to stabilise visual field progression 3) Failure to stabilize optic disc damage progression
Secondary outcome(s)	1) Period (days/weeks) taken to complete healing 2) Percentage/proportion/ length/area of ulcer healed after a period (days or months) 3) Number of ulcer recurrence after healing 4) Interval of time between healing and recurrence 5) Level of visual acuity 6) Pain assessment at intervals of treatment period using 0 to 100 score, visual analogue score or any form of pain measurement adopted by the trialists 7) Analgesia use 8) Use of topical cycloplegics 9) Measure of compliance to treatment 10) Other symptoms, for example, photophobia. 11) Severe adverse effects of interventions 12) Minor adverse effects of interventions 13) Quality of life measures 14) Economic data	1) Necessity of adding or changing medical therapeutic regimen in consequence of uncontrolled IOP and visual field or optic disc damage progression 2) Adverse effects (severe, minor) including: • intraocular pressure spikes • uveitis • ciclitis • hyphema • goniosinechiae formation • corneal edema • persistent intraocular pressure elevation • loss of vision (central island) • syncope 3) Quality of life measures 4) Economic data
Number of primary outcomes	1	3
Number of secondary outcomes	Unclear: >14	Unclear: >4

identified control groups or with each other. For example, to ameliorate postoperative pain following emergence from anesthesia, preventative treatments may be started during the immediate operative period. These treatments include: (1) the administration of opioids, local anesthetics or both into the epidural space, (2) the infiltration of local anesthetics directly into or near the surgical wound, (3) the systemic administration of drugs that are *N*-Methyl-D-Aspartic Acid antagonists, (4) the systemic administration of non steroidal anti-inflammatory drugs, and (5) the systemic administration of opioids.

It has been speculated that administering an analgesic before the surgical incision will establish a "preemptive" analgesic effect that will outlast the actual residence time of the drugs in the body. Several trials have compared the postoperative pain experience of patients receiving the same treatment immediately before (early application) vs. immediately after surgery (late application). Ong et al. [15] found a total of 66 randomized controlled trials comparing early vs. late application for the five types of intervention described above. Three primary outcomes were obtained from the trials: an ordinal postoperative pain score during the first 1–2 days, the amount of supplemental analgesic administered postoperatively, and the time to the first administration of supplemental analgesic. Meta-analyses were performed using standardized mean differences comparing early vs. late application for each of the five interventions for each of these three primary outcomes. In the resulting 15 comparisons, seven favored early, one favored late, and seven comparisons were not statistically significant (Table 2). No multiplicity adjustment for the multiple early vs. late comparisons was made.

One strategy for avoiding multiple testing is to specify, a priori, which comparisons are of primary interest in the protocol for the review. If individual patient data are available, appropriate methods to deal with multiple intervention groups are given by known multiple comparison procedures within the framework of analysis of variance [16]. If only summary statistics of studies are available, the application of multivariate meta-analysis can be considered [17, 18]. However, appropriate methods to compare multiple groups with adjustment for multiple testing in the framework of multivariate meta-analysis have yet to be developed.

3.3. Multiple time points

A multiplicity problem due to multiple time points can principally arise in three ways: (1) the outcomes are measured at specific, but different points in time for each participant, (2) the participants have different durations of follow up, and (3) the (accumulating) data are analyzed at different time points (interim analyses) for the monitoring of trials. The latter can also be a problem in cumulative meta-analyses (see Section 3.6).

Patient outcomes are often measured at different time points for pragmatic reasons. In studies with ongoing follow up, where the amount of follow up for each patient

Table 2
Summary statistics from meta-analyses comparing presurgical and postsurgical initiation of five analgesic interventions regarding three outcomes [15]

Intervention	Pain score SMD (95% CI)	Supplemental analgesic SMD (95% CI)	Time to first analgesic SMD (95% CI)
Epidural analgesia	0.25 (0.10, 0.41)*; $P < 0.01$	0.58 (0.42, 0.74)*; $P < 0.01$	0.31 (0.10, 0.52)*; $P < 0.01$
Local anesthesia	0.10 (−0.07, 0.27); $P = 0.26$	0.44 (0.23, 0.65)*; $P < 0.01$	0.44 (0.21, 0.68)*; $P < 0.01$
NMDA antagonist	0.00 (−0.19, 0.20); $P = 0.97$	0.17 (−0.02, 0.37); $P = 0.09$	0.12 (−0.13, 0.37); $P = 0.34$
NSAID	0.14 (−0.02, 0.30); $P = 0.09$	0.48 (0.31, 0.64)*; $P < 0.01$	0.68 (0.44, 0.91)*; $P < 0.01$
Opioids	−0.24 (−0.46, −0.01)**; $P = 0.04$	0.23 (−0.06, 0.52); $P = 0.12$	−0.34 (−0.81, 0.13); $P = 0.16$

Standardized mean differences (SMDs) with 95% confidence intervals (CIs) were estimated by using fixed effect models. A SMD > 0 favors a presurgery intervention (indicated by *); a SMD < 0 favors a post surgery intervention (indicated by**).

Epidural analgesia: the epidural administration of opioids and/or local anesthetics; Local anesthesia: infiltration of the surgical wound with local anesthetics; NMDA antagonist: systemic administration of *N*-Methyl-D-Aspartic Acid antagonist; NSAID: systemic administration of nonsteroidal anti-inflammatory drug; Opioids: systemic administration of opioids.

and for each trial might differ, this is an inherent problem. It is resolved if all patients have the same duration of follow up but this is unlikely and a common problem in systematic reviews is that follow-up times differ among the included studies. The situation is further complicated by patient attrition, which means that the analyzed populations differ at the time points, even within a single trial.

The multiplicity problem can be solved in a single trial by choosing one single time point for the primary analysis. However, for chronic diseases it is of special interest to examine the course of the patients over time. So, it may be difficult to choose one single time point in a trial as the main one. For example, in clinical trials of minimally invasive procedures for the treatment of benign prostatic hyperplasia, the patient's symptoms might be assessed after 3, 6, and 12 months, but also for several years after the first without addressing the problem of multiple time points [19].

Another possibility to overcome the multiplicity problem without the necessity to choose a primary time point is to use a summary effect measure over time (e.g., via repeated measures analysis of variance for continuous outcomes or Cox regression for time-to-event data). In systematic reviews, such solutions are applicable if sufficient data are available from the included studies [20]. In most cases, however, a satisfactory statistical solution requires the availability of individual patient data from the studies, but it can also be achieved if the trialists supply the results of the same standard analysis for their study [21].

3.4. Multiple effect measures

The multiplicity that might arise in a systematic review because of the variety of outcomes is compounded further by the variety of ways to analyze these to obtain an estimate of the difference between the intervention and the control group. In some reviews, as in the example given below, a variety of different methods might be tried for the meta-analysis to assess the robustness of the findings to the choice of analysis technique. This could be done in a series of sensitivity analyses. However, a multiplicity problem arises if the reviewer calculates the overall effect estimate in a variety of ways and then focuses on the one that gives them the finding that they wish to emphasize. If a meta-analysis is done, this multiplicity arises at the level of the overall effect estimate. If the results of each study are kept separate, with no combination of these in a meta-analysis, it arises at the study level.

If a meta-analysis uses continuous data, the usual effect measure is the mean difference. If the included studies measure the same outcome in different ways, for example, by using a variety of scales to measure anxiety, the effect measure will need to be standardized. The reviewers might also choose other options such as dichotomizing continuous data into ''high'' and ''low.'' In meta-analyses of binary data, typical effect measures are given by the odds ratio, risk ratio, and absolute risk difference. Multiplicity becomes important and problematic if there is an important difference in the conclusions of the review depending on which of these effect measures is used. In such circumstances, it is especially important that the appropriateness of the effect measure for the type of data being meta-analyzed is considered [22].

For example, a systematic review of the effects on deep vein thrombosis of wearing special compression socks on long haul flights illustrates the potential problems of multiplicity in the choice of effect measures [23]. The reviewers found nine trials in which passengers were randomized to either wear, or not wear, the socks during a flight. A total of nearly 3,000 people were randomized, and the original intention of the reviewers was that their meta-analyses would use the odds ratio calculation available in The Cochrane Collaboration's software, RevMan 4.2. However, events in the trials were rare. In total, only 50 asymptomatic deep vein thromboses were identified. There were no asymptotic deep vein thromboses in three of the nine trials and, typically, only a few percent of the passengers in the other trials had this outcome. The default method used to calculate an odds ratio in RevMan is the Mantel-Haenszel method and this produced an odds ratio of 0.10 (95% CI 0.04–0.25, $P < 0.00001$). However, the RevMan analysis adds a continuity correction of 0.5 if there are no events in one group, which might be too high for the data in this review [24]. The reviewers performed sensitivity analyses using the Mantel-Haenszel method with a variety of small continuity corrections, logistic regression, and the Peto

odds ratio method. This showed that the choice of analysis technique made little important difference to the overall conclusions. The recalculated odds ratios using the Mantel–Haenszel method converged to a steady level as the continuity correction was diminished to zero, which was identical to the result from logistic regression. The result using the Peto method was also similar [23]. Thus, using multiple effect measures for this review revealed the stability of its conclusions (i.e., compression socks greatly reduce the risk of asymptomatic deep vein thromboses in long-haul airline passengers), because the results were consistent across the different analyses. However, if the results had been different in important ways, it would have been difficult to interpret and to determine which effect measure provided the most reliable estimate. It would be inappropriate to perform the analysis in several ways and then to present only that one yielding a significant result.

3.5. Subgroup analyses

Subgroup analyses are separate analyses of a subset of the trials within a meta-analysis or of a subset of participants in a single study or the meta-analysis as a whole. They are usually done to investigate whether treatment effects vary according to characteristics of the populations, interventions or studies. In systematic reviews, two different sorts of subgroup analysis can be recognized: (1) whole trials are classified into subgroups based on characteristics of the trials or their participants (e.g., a comparison of placebo-controlled vs. non–placebo-controlled trials), (2) participants are classified into subgroups, so that each trial might contributes data to several or all subgroups (e.g., a comparison of the effects in men and women).

The potential of subgroup analyses to produce misleading results has been demonstrated many times in randomized trials and systematic reviews [25–27]. Subgroup analyses can increase the number of analyses enormously. This causes major problems with multiple comparisons, and thus, greatly increases the potential for spurious effects and false conclusions. For example, if a review performs naive subgroup analyses based on five different patient characteristics, each splitting the population into two subgroups, with 10 outcomes being analyzed, this would add 100 extra analyses to the review. The possible number of extra comparisons would increase even further if some characteristics generate more than two subgroups, or if the subgroup analyses were repeated for each of several different comparisons.

The systematic review of antenatal corticosteroids for accelerating fetal lung maturation in women at risk of preterm birth [28], which provided the inspiration for The Cochrane Collaboration's logo, is a good example of the effects of multiple subgroup analyses on the number of comparisons performed. For one of the main outcomes (respiratory distress syndrome), there were 10 different subgroup analyses, in which a separate analysis was performed for each subgroup. These were single vs. multiple pregnancy (two groups), gestational age at birth (seven nonmutually exclusive subgroups), timing of birth after randomization (four nonmutually exclusive groups), whether or not there was prelabour rupture of membranes (two groups), prolonged membrane rupture (two groups), whether the pregnancy was complicated by hypertension (one group), corticosteroid drug used (two groups), decade of trial (three groups), gestational age at randomization (six nonmutually exclusive groups), and single or repeated courses (two groups). In total, 32 analyses were performed for the outcome respiratory distress syndrome alone: one main analysis and 31 analyses of subgroups. With similar numbers of analyses being performed for other outcomes, it is likely that some conventionally significant results arose by chance alone.

Erroneous results from subgroup analyses may result not only from the number of analyses performed, but also from the way they are performed. Subgroup analyses compare the effects of an intervention among different sorts of trials or different sorts of participants. Formal statistical methods using an appropriate test for interaction are therefore needed to assess whether there is any evidence of a difference in treatment effect between the subgroups. If separate analyses are conducted for each subgroup and the results compared, it is easy to find a statistically significant result in one subgroup but not in another. This may be incorrectly interpreted as evidence that the treatment is effective in only one subgroup. In fact, even if there is no difference in the treatment effect estimates among the subgroups, it is possible that only one will be statistically significant because that subgroup contains the most trials or participants. Judging subgroup differences solely by the results of significance tests within subgroups is therefore invalid and an appropriate interaction test should always be used for subgroup analyses [29].

The number of comparisons generated by subgroup analyses can be reduced by keeping the number of subgroups analyzed to a minimum, and ensuring that each has a clear rationale, such as a plausible biological mechanism, or previous research suggesting a difference between different types of patient. Restriction of subgroup analyses to a subset of the most important outcome variables is also recommended, because it helps to avoid unnecessary inflation of the number of analyses.

Post hoc subgroup analyses are exploratory in nature and can only be used for generating new hypotheses, which require subsequent independent investigation. In general, they should not be used for the main conclusions of a review. If the goal of subgroup analyses in the context of systematic reviews is to investigate heterogeneity, which is essential to justify that the average effect estimate from a meta-analysis of several studies is sensible, no adjustments for multiple comparisons are required, because the analyses belong to different questions. However, if subgroup analyses are performed to investigate the main

objective of the systematic review, multiplicity should be taken into account. In the case of one primary outcome and one primary subgroup-generating variable, multiplicity is avoided because only one interaction test is required. In the case of research questions leading to several outcomes or several subgroup generating variables, appropriate adjustments for multiple comparisons are desirable (see Section 3.1).

3.6. Multiple looks at accumulating data

Meta-analyses in systematic reviews are often done as if the results of the included trials had not been previously summarized to obtain an overall estimate of the treatment effect. However, systematic reviews are commonly updated as new trials are published, that is, repeated analyses are performed on accumulating data. For example, Cochrane reviews should be updated periodically, at least every second year [30]. Furthermore, to strengthen inferences some reviewers might choose to conduct cumulative meta-analyses (which involve repeated calculations of the overall statistics) after every publication time point of the included trials [31, 32]. In both cases, subsets of the included trials are analyzed multiple times using the conventional significance level of $\alpha = 5\%$. Such multiple looks on accumulating evidence in meta-analysis adds another dimension of multiplicity to systematic reviews, namely *repeated significance testing*. The situation is comparable to interim analyses of clinical trials, where it has been recognized for a long time that adequate repeated analyses of accumulating data must include adjustments for multiple significance testing [33].

A wide range of statistical methods is available for multiplicity adjustments in accumulating data from clinical trials [34]. Pouge and Yusuf [35] proposed to use classical O'Brien–Fleming monitoring boundaries for use in cumulative meta-analyses. Devereaux et al. [36] applied monitoring boundaries to quantify the degree of evidence available for the use of perioperative beta-blockers in noncardiac surgery. Presumably, authors will need to keep updating their systematic reviews until there is a sufficient amount of evidence available to draw firm conclusions. This situation is analogous to continuing clinical trials until either the number of included patients exceeds a predetermined sample size or the resulting cumulative test statistic crosses a stopping boundary. Therefore, it has been argued that the standards for significance testing in meta-analyses should be no less rigorous than those of a single trial [35, 36]. Wetterslev et al. [37] derived a correction factor to adjust the required meta-analysis sample size for heterogeneity and proposed that this method should be used for prospective planning and testing in future cumulative meta-analyses. However, the available methods for multiplicity adjustments in accumulating data are based on assumptions that are hardly met in any meta-analysis. The desired error rates may therefore not be achieved when the available methods are adapted to meta-analysis [38]. Lan et al. [39] and Hu et al. [40] proposed methods to control the inflation of the test statistic accounting for multiple testing, heterogeneity, and the unpredictable nature of information from trials for continuous and binary outcomes.

Cumulative meta-analyses have been used for the retrospective purpose of identifying the point in time where a treatment effect reached the level of statistical significance for the first time [31]. For example, it was concluded that rofecoxib should have been withdrawn from the market several years before 2004, because evidence for an increased cardiovascular risk of this drug was available as early as 2000 [41]. Some authors argue that accumulating meta-analyses are best interpreted in a Bayesian framework, and that there is no need to adjust for multiplicity in cumulative meta-analyses [32, 42].

The relevance of adjusting for multiplicity in cumulative meta-analysis data depends on whether the underlying systematic review is for descriptive or decision-making purposes. If cumulative meta-analysis is used as retrospective tool for descriptive purposes, adjustments for multiple testing are not required. If, however, a final decision is based upon prospectively planned accumulating evidence, an adjustment for multiple testing in meta-analyses seems to be reasonable. The recently proposed approaches in this field deserve attention in future research [37, 39, 40].

4. Guidelines for dealing with multiplicity in systematic reviews

The following general guidance should help to reduce the problems caused by multiple comparisons in systematic reviews.

4.1. General issues

- Interpret any results that were not based upon a priori formalized hypotheses cautiously, even when they are "statistically significant." Such findings should not be regarded as proof of a hypothesis but, rather, as a means of generating hypotheses.
- Define which analyses belong to one objective, for which the type 1 error should be under control. It is possible to have several objectives and to use the significance level $\alpha = 5\%$ for each objective, but if the results regarding several objectives are combined into one final conclusion, the type 1 error for the final conclusion is no longer under control. If one objective leads to multiple comparisons, the use of a multiple testing procedure is required to control the MEER. The use of adequate adjustments for multiple comparisons will lead to stronger evidence.
- A number of multiple comparison procedures (for multiple significance testing as well as for multiple CIs) have been developed for use in single trials.

These methods can also be applied in meta-analyses when individual patient data are available. Some of these methods can also be used in meta-analyses when only aggregated data are available, for example, general multiple comparison procedures based upon *P*-values [4]. However, more research is required to develop appropriate multiple comparison procedures for use in meta-analyses with aggregated data (e.g., in the framework of multivariate meta-analysis).
- At the very least, authors and users of systematic reviews should take the problems of multiplicity into account when they interpret findings based upon multiple comparisons.

4.2. Multiple outcomes

- State which outcomes are of particular interest in the protocol of the review (the fewer the better). It is helpful to classify the outcomes into primary outcomes, which will be used for drawing the review's main conclusions, and secondary outcomes, which will not contribute to the main conclusions. Consider whether it is necessary to perform meta-analyses for secondary outcomes, or whether it would be better simply to describe these results or to tabulate them without meta-analyses.
- Although it is recommended that Cochrane reviews should seek to include all outcomes that are likely to be important to users of the review, the overall strength of evidence will be less if there are multiple comparisons without corresponding adjustments.

4.3. Multiple groups

- State which groups and comparisons are of particular interest in the protocol of the review (the fewer the better).
- Describe in detail the main hypotheses. If there are more than two treatment groups, there are a lot of possibilities (e.g., any difference among the groups, difference between a number of groups and one control group, differences concerning all pairwise comparisons).

4.4. Multiple time points

- Present a summary effect over all time points or choose one primary time point that is the most appropriate if there is a choice of time points for an outcome.
- Avoid multiple testing of the effect at each of the time points.

4.5. Multiple effect measures

- Choose an effect measure when the systematic review is being planned, based on its mathematical appropriateness for the type of data to be analyzed. Describe in a transparent way any modification or substitution of the originally intended effect measure if it is found to be inappropriate or suboptimal during data analysis.
- Be cautious about dichotomizing continuous data into "high" and "low" values, especially if there is no natural or well accepted threshold for the "high/low" division.

4.6. Subgroup analyses

- Prespecify subgroup analyses to be conducted and keep them to a minimum. Ensure that each has a clear rationale, which should be explained in the background to the review.
- Restrict subgroup analyses to a small subset of outcomes (in most cases the primary outcomes).
- Interpret the results of subgroup analyses by using an appropriate heterogeneity or interaction test.
- Do not adjust for multiple testing concerning subgroup analyses performed to investigate heterogeneity and to justify the subsequent main analysis.
- Interpret subgroup analyses cautiously and do not base the review's conclusion on exploratory subgroup analyses. Seek independent evidence to justify any claims that the effects of the interventions vary in important ways among subgroups.

4.7. Accumulating data

- Incorporate sample size considerations and the expected degree of heterogeneity for the primary effect measure(s) in prospectively planned cumulative meta-analyses. Adjust the threshold for statistical significance or the inflation of the test statistic to account for future multiple looks as data accumulate if the required sample size is not yet reached and future review updates are planned.
- Do not adjust for multiple testing in cumulative meta-analyses used as a retrospective tool for descriptive purposes.

5. Conclusion

We pointed out that multiple comparisons represent a problem in systematic reviews concerning several described situations. The main problem is the same, regardless of whether multiple significance tests are performed

or multiple CIs are calculated. The risk of finding spurious associations or effects increases with the number of comparisons so that ''evidence'' described in the review might be based on chance alone rather than a true effect. Solutions to deal with multiplicity in single trials can sometimes, but not always, be applied in systematic reviews. A priori planning of multiple comparison procedures in systematic reviews is frequently difficult or even impossible because the data situation is not known in advance. There are some promising meta-analytic approaches (e.g., multivariate meta-analysis) that offer possibilities to account for multiplicity. However, more research is required to develop multiple comparison procedures that are feasible in systematic reviews.

In conclusion, there is no simple or completely satisfactory solution to the problem of multiple comparisons in systematic reviews. It is, however, an issue that requires recognition. Authors and users of reviews need to be careful about multiplicity when presenting, interpreting and using reviews that contain or are based on numerous statistical analyses. We have offered suggestions for dealing with this problem, and recommend that the multiplicity issue requires acknowledgment in any review with multiple comparisons.

References

[1] Bauer P. Multiple testing in clinical trials. Stat Med 1991;10:871–90.

[2] Ottenbacher KJ. Quantitative evaluation of multiplicity in epidemiology and public health research. Am J Epidemiol 1998;147:615–9.

[3] Biester K, Lange S. The multiplicity problem in systematic reviews [Abstract]. *XIII Cochrane Colloquium, Melbourne, 22–26 October 2005*. Program and Abstracts; 2005: 153. Available at http://www.cochrane.org/colloquia/abstracts/melbourne/P-155.htm. Accessed May 13, 2008.

[4] Bender R, Lange S. Adjusting for multiple testing—when and how? J Clin Epidemiol 2001;54:343–9.

[5] Cook RJ, Dunnett CW. Multiple comparisons. In: Armitage P, Colton T, editors. Encyclopedia of biostatistics. Chichester: Wiley; 2005. p. 3383–93.

[6] Dmitrienko A, Hsu JC. Multiple testing in clinical trials. In: Kotz S, Read CB, Balakrishnan N, Vidakovic B, editors. Encyclopedia of statistical sciences. Hoboken, New Jersey: Wiley; 2006. p. 5111–7.

[7] Chen T, Hoppe FM. Simultaneous confidence intervals. In: Armitage P, Colton T, editors. Encyclopedia of biostatistics. Chichester: Wiley; 2005. p. 4953–5.

[8] Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995;310:170.

[9] Rabiu M, Alhassan MB, Agbabiaka IO. Interventions for Mooren's ulcer (Protocol). Cochrane Database Syst Rev 2006;3. CD006131.

[10] Rolim de Moura C, Paranhos A Jr. Laser trabeculoplasty for open angle glaucoma (Protocol). Cochrane Database Syst Rev 2002;4. CD003919.

[11] van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med 2002;21:589–624.

[12] Arends LR, Voko Z, Stijnen T. Combining multiple outcome measures in a meta-analysis: an application. Stat Med 2003;22:1335–53.

[13] Riley RD, Abrams KR, Sutton AJ, Lambert PC, Thompson JR. Bivariate random-effects meta-analysis and the estimation of between-study correlation. BMC Med Res Methodol 2007;7:3.

[14] Ishak KJ, Platt RW, Joseph L, Hanley JA. Impact of approximating or ignoring within-study covariances in multivariate meta-analyses. Stat Med 2008;27:670–86.

[15] Ong CK, Lirk P, Seymour RA, Jenkins BJ. The efficacy of preemptive analgesia for acute postoperative pain management: a meta-analysis. Anesth Analg 2005;100:757–73.

[16] Hsu JC. Multiple comparisons: Theory and methods. Boca Raton, FL: Chapman & Hall; 1996.

[17] Hasselblad V. Meta-analysis of multitreatment studies. Med Decis Making 1998;18:37–43.

[18] Riley RD, Abrams KR, Lambert PC, Sutton AJ, Thompson JR. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. Stat Med 2007;26:78–97.

[19] Bouza C, Lopez T, Magro A, Navalpotro L, Amate JM. Systematic review and meta-analysis of transurethral needle ablation in symptomatic benign prostatic hyperplasia. BMC Urol 2006;6:14.

[20] Parmar MK, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. Stat Med 1998;17:2815–34.

[21] Dear KBG. Iterative generalized least squares for meta-analysis of survival data at multiple times. Biometrics 1994;50:989–1002.

[22] Deeks JJ, Altman DG. Effect measures for meta-analysis of trials with binary outcomes. In: Egger M, Davey Smith G, Altman DG, editors. Systematic reviews in health care: Meta-analysis in context. London: BMJ Books; 2001. p. 313–35.

[23] Clarke M, Hopewell S, Juszczak E, Eisinga A, Kjeldstrom M. Compression stockings for preventing deep vein thrombosis in airline passengers. Cochrane Database Syst Rev 2006;2. CD004002.

[24] Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Stat Med 2004;23:1351–75.

[25] Davey Smith G, Egger M. Going beyond the grand mean: subgroup analysis in meta-analysis of randomised trials. In: Egger M, Davey Smith G, Altman DG, editors. Systematic reviews in health care: Meta-analysis in context. London: BMJ Books; 2001. p. 143–56.

[26] Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Stat Med 2002;21:2917–30.

[27] Austin PC, Mamdani MM, Juurlink DN, Hux JE. Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. J Clin Epidemiol 2006;59:964–9.

[28] Roberts D, Dalziel S. Antenatal corticosteroids for accelerating fetal lung maturation for women at risk of preterm birth. Cochrane Database Syst Rev 2006;3. CD004454.

[29] Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. Health Technol Assess 2001;5:1–56.

[30] Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions 4.2.6 [updated September 2006]. The Cochrane Library, Issue 4, 2006. Chichester, UK: Wiley; 2006.

[31] Lau J, Antman EM, Jimenez-Silva J, Kupelnick B, Mosteller F, Chalmers TC. Cumulative meta-analysis of therapeutic trials for myocardial infarction. N Engl J Med 1992;327:248–54.

[32] Lau J, Schmid CH, Chalmers TC. Cumulative meta-analysis of clinical trials builds evidence for exemplary medical care. J Clin Epidemiol 1995;48:45–57.

[33] McPherson K. Statistics: The problem of examining accumulating data more than once. N Engl J Med 1974;290:501–2.

[34] O'Brien PC. Data and safety monitoring. In: Armitage P, Colton T, editors. Encyclopedia of biostatistics. Chichester: Wiley; 2005. p. 1362–71.

[35] Pogue JM, Yusuf S. Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. Control Clin Trials 1997;18:580–93.

[36] Devereaux PJ, Beattie WS, Choi PT, Badner NH, Guyatt GH, Villar JC, et al. How strong is the evidence for the use of perioperative beta

blockers in non-cardiac surgery? Systematic review and meta-analysis of randomised controlled trials. BMJ 2005;331:313–21.

[37] Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. J Clin Epidemiol 2008;61:64–75.

[38] Whitehead A. A prospectively planned cumulative meta-analysis applied to a series of concurrent clinical trials. Stat Med 1997;16:2901–13.

[39] Lan KKG, Hu M, Cappalleri JC. Applying the law of iterated logarithm to cumulative meta-analysis of a continuous endpoint. Statistica Sinica 2003;13:1135–45.

[40] Hu M, Cappelleri JC, Lan KK. Applying the law of iterated logarithm to control type I error in cumulative meta-analysis of binary outcomes. Clin Trials 2007;4:329–40.

[41] Jüni P, Nartey L, Reichenbach S, Sterchi R, Dieppe PA, Egger M. Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. Lancet 2004;364:2021–9.

[42] Ioannidis JPA, Contopoulos-Ioannidis DG, Lau J. Recursive cumulative meta-analysis: a diagnostic for the evolution of total randomized evidence from group and individual patient data. J Clin Epidemiol 1999;52:281–91.