Standardised mean difference in meta-analyses - How reliable is it in practice?

Britta Tendal

PhD thesis

Faculty of Health Sciences
University of Copenhagen

Standardised mean difference in meta-analyses - How reliable is it in practice?

Britta Tendal

This thesis is accepted for defense by the
Faculty of Health Sciences, University of Copenhagen
26th July 2010
The defense takes place 26th August 2010
at Medicinsk Museion , Bredgade 62, Copenhagen

Supervisors:
Peter C. Gøtzsche
Peter Jüni
Julian Higgins

Assessment Committee:
Per Kragh Andersen
Douglas G. Altman
Gerd Antes

Britta Tendal
The Nordic Cochrane Centre
Rigshospitalet, Department 3343
Blegdamsvej 9
DK-2100 Copenhagen
Denmark
Phone: +45 35 45 71 46
Fax: +45 35 45 70 07
E-mail: bt@cochrane.dk
Web: www.cochrane.dk

Table of contents

Preface and acknowledgements 2
Structure of the thesis 2
Acknowledgements 2
Financial support 2
Original papers 3
Abstract 4
Danish summary 5
Introduction 6
Systematic reviews 6
Known challenges with systematic reviews 7
The standardised mean difference 7
New challenges with systematic reviews 8
Objectives 9
Methods & results 10
Paper 1 10
Discussion of paper 1 19
Paper 2 20
Discussion of paper 2 28
Paper 3 30
Discussion of paper 3 47
Conclusions 49
Future research 49
Reference list 51

PREFACE AND ACKNOWLEDGEMENTS

STRUCTURE OF THE THESIS

The structure of this thesis follows the guidelines from the Faculty of Health Science, University of Copenhagen. First I will give a brief introduction to the context of this thesis and present the objectives. The next part of the thesis consists of three articles, each followed by a discussion including a critical appraisal of the methods and conclusions and a comparison with published results of other researchers. Finally, I present an overall conclusion and suggestions for future research.

ACKNOWLEDGEMENTS

It is a pleasure to thank those who made this thesis possible. First I owe my deepest gratitude to my main supervisor Peter C. Gøtzsche, thank you for opening the door to research for me and for always having time for questions. Second I would like to thank my two other supervisors Peter Jüni and Julian Higgins. They have both contributed with indispensable support and have like my main supervisor generously poured of their knowledge and experience.

I would also like to show my gratitude to all my co-authors, I would like to thank them all for their collaboration on the papers and for allowing me to lock them up in a hotel for a week. A special thanks goes to Eveline Nüesch for valuable discussions and productive teamwork.

I am indebted to all my colleagues for their support and for making my three years of PhD studies an enjoyable time. Especially I would like to thank Asbjørn Hróbjartsson and Karsten Juhl Jørgensen for interesting discussions about science and constructive criticism. I would also like to thank Marian Pandal, Frihild Askham and Jannie Hedegaard for secretarial assistance. Furthermore I would like to thank Rasmus Moustgaard for assisting with searches in the Cochrane Database of Systematic Reviews and Jacob Riis for all his help with the layout of my posters. I am grateful to Georgie Imberger for reading various drafts and giving me many helpful suggestions.

I would also like to thank Arthur Tudsborg for the illustrations used in this thesis.

Finally I am very grateful to my husband Ulf and my children Elisabeth and Robin for their love and support.

FINANCIAL SUPPORT

I would like to thank IMK Charitable Fund and The Nordic Cochrane Centre, which funded this PhD.

ORIGINAL PAPERS

The thesis is based on the following papers:

Gøtzsche PC, Hróbjartsson A, Maric K, Tendal B. Data extraction errors in meta-analyses that use standardized mean differences. JAMA. 2007 Jul 25;298(4):430-7.

Tendal B, Higgins JP, Jüni P, Hróbjartsson A, Trelle S, Nüesch E, Wandel S, Jørgensen AW, Gesser K, Ilsøe-Kristensen S, Gøtzsche PC. Disagreements in meta-analyses using outcomes measured on continuous or rating scales: observer agreement study. BMJ. 2009 Aug 13;339:b3128.

Tendal B, Higgins JP, Jüni P, Nüesch E, Gøtzsche PC. Multiplicity of data in trial reports creates an important challenge for the reliability of meta-analyses: an empirical study. Manuscript submitted.

ABSTRACT

The use of the standardised mean difference (SMD) is common in meta-analyses, as it allows outcomes of a similar nature, but measured on different scales, to be combined. The application of SMDs, compared with that of the raw mean difference, can be complex. Despite this complexity, there have been few studies of the reliability of this effect measure in practice.

The aims of this PhD were to investigate the difficulties that may arise when researchers use SMD as an effect measure and to determine the scope for reviewer bias. Three studies were undertaken in order to fulfil these aims. In the first study, we evaluated the reproducibility of meta-analyses using SMDs. In the second study, we determined the observer variation when extracting data for the computation of SMDs. In the third study, we investigated the range of SMDs that could be calculated based on the same outcomes from the same trials.

The results from the three studies demonstrate that data extraction is prone to error, which can negate or even reverse the findings of studies. Disagreements were common and often larger than the effect of commonly used treatments and multiplicity of data was frequent and could impact importantly on meta-analytical results.

The conclusion is that readers should be aware that meta-analyses using SMDs may have included incorrect or selectively extracted data. Readers should seek assurances that data collection methods were rigorous and clearly predefined. Reliability of meta-analyses using SMDs might be improved by having more detailed review protocols, more than one observer, and investigators with statistical expertise, but this has to be confirmed by future research.

DANISH SUMMARY

Den standardiserede gennemsnitlige forskel (SMD) er alment brugt til meta-analyser, da metoden tillader lignende udfald målt på forskellige skalaer at blive kombineret. På trods af at metoden er mere indviklet at anvende end for eksempel gennemsnitlige forskelle, er der få studier, som klarlægger, hvor pålidelig metoden er i praksis.

Formålene med denne PhD var at undersøge, hvilke udfordringer forskere støder på, når de anvender metoden samt at fastslå hvilke muligheder for bias, der er forbundet med metoden. Vi udførte tre studier at opfylde disse formål. I det første studie vurderede vi reproducerbarheden af publicerede meta-analyser, i det andet studie fastslog vi omfanget af uenighed blandt observatører i forbindelse med ekstraktion af data og i det tredje studie undersøgte vi hvilke forskellige SMDs, der kunne beregnes baseret på de samme effektmål og de samme forsøg.

Resultaterne fra de tre undersøgelser viste, at ekstraktion af data kan medføre fejl i et omfang, som kan neutralisere eller endog vende resultaterne af undersøgelser, at uenigheder var almindelige og ofte større end effekten af almindeligt anvendte behandlinger, og at de samme effektmål og forsøg kunne give flere forskellige resultater for hvert forsøg hvilket påvirker de meta-analytiske resultater.

Vores konklusion er, at læsere bør tolke meta-analyser med SMD'er med forsigtighed, da de kan være baseret på ukorrekte eller selekterede data. Læsere bør være opmærksomme på, hvorvidt metoderne til ekstrationen af data var stringente og beskrevet på forhånd. Meta-analyser kan sandsynligvis forbedres ved hjælp af mere detaljerede protokoller, mere end en observatør og adgang til statistisk ekspertise, dette skal dog efterprøves i fremtidige studier.

INTRODUCTION

Decisions regarding health care interventions should be evidence based. Medical interventions have the potential to be both beneficial and harmful. The goal of medical research is to define the benefits and harms reliably, allowing health professionals and patients to make informed decisions about healthcare interventions. Health care is also delivered within limited financial resources, stressing the need to identify efficacious treatments. Such assessments of health care interventions necessitate a ranking of evidence. When examining the effects of health care interventions, the recommended hierarchy is based on the degree of bias connected with the study designs. There are many different ways of ordering different study design into a hierarchy. Usually, observational studies such as cross-sectional, cohort and case-control studies are ranked below randomised trials. This ranking is because observational study designs are more prone to bias than randomised studies (1). Randomised trials have the advantage of having a control group, which allows the natural progression of the disease to be taken into account when evaluating the effect of a treatment. Another important aspect is the randomisation, which balances known as well as unknown prognostic factors between the intervention groups (2). Systematic reviews of RCTs, are considered to be the highest ranking quality of evidence since they combine the advantages of RCTs with a higher degree of precision (3).

SYSTEMATIC REVIEWS

Systematic reviews are becoming widespread. In a study by Moher et al., it was estimated that about 2,500 new systematic reviews are indexed annually on Medline (4). This increase highlights the need to investigate and assess the challenges connected with this research design.

A systematic review is not simply a summary of existing evidence; it is an investigation in its own right. The aim of a review is to define a specific research question and answer it with a systematic and robust technique. A systematic review should have a protocol, which clearly defines the research question and pre-specifies the methods of the review including the search strategy, which studies to in- and exclude, which data to extract and which analyses to perform. The conduct of the review then follows that protocol with any necessary changes clearly described and explained in the final publication. There is a comprehensive search for all relevant studies, a selection process based on the pre-specified inclusion and exclusion criteria, an assessment of the validity of the included studies and extraction of the relevant data. Analyses based on the extracted data are performed, these can be narrative, such as a structured summary and discussion of the studies' characteristics and findings, or quantitative, that is involving statistical analyses which can either be narrative or quantitative. The process of combining results quantitatively from different studies is called a meta-analysis. The results of the analyses along with the assessments of validity are interpreted and reported (3;5;6).

If a systematic review is well performed and reported, its findings are a valuable base for decision-making. As with all research, there are advantages and disadvantages to this technique of summarising evidence. In order to make informed healthcare decisions, it is important to understand the strengths

and weakness of conclusions made in a systematic review, and to use the knowledge with these in mind. One of the strengths of a systematic review with meta-analyses is that the precision of the estimates is likely to be increased, allowing effects to be detected that might be missed in single trials. Another advantage is that the consistency of the results across trials can be investigated and it may be possible to detect bias. If trials give conflicting evidence, the degree of inconsistency can be quantified (3).

KNOWN CHALLENGES WITH SYSTEMATIC REVIEWS

The process of combining data from multiple trials has created new challenges, some of which may compromise the reliability of the conclusion. Sometimes, studies that are too diverse are combined, giving results that are difficult to interpret. The issue of diversity is an important one and meta-analysts should consider two causes of diversity: methodological and clinical. Methodological diversity is caused by differences in study design and thereby different risks of bias. Clinical diversity is caused by differences in participants, interventions and outcomes across the studies (3).

The results of a systematic review can only be as robust as the included trials. While meta-analyses can improve precision by virtue of having higher power than the individual trials, the validity of the conclusion depends on the validity of the included trials. If a systematic review contains studies with a high risk of bias, the meta-analytical result will also contain a high risk of bias, despite any enhanced precision of combined effect measures. Clearly, it is important to include an assessment of risk of bias in the discussion of the results and the conclusion.

An additional challenge is publication bias. Not all studies are published and if there are any systematic differences in regard to which studies are published, this difference can lead to bias in the results of meta-analyses (7;8). The same is true for selective outcome reporting were only some of the outcomes or analyses are presented (9-11). It is often difficult to know how to manage missing information in medical research. The first step is to identify its presence. Performing a meta-analysis can give an overview of the existing studies and make it possible to see patterns that might indicate missing studies (3).

THE STANDARDISED MEAN DIFFERENCE

In this thesis, I have chosen to focus on one specific effect measure. The selected measure is the standardised mean difference. The use of standardised mean differences (SMDs) is common in meta-analyses, because it allows outcomes of a similar nature to be combined. Sometimes researchers investigate outcomes like pain, which can be measured on different scales. In order to combine these results into one overall estimate, they need to standardise the scales. For example, combining a measurement on a 7 point ranking scale with one on a 100 mm visual analogue scale by taking the average would be meaningless. These two measurements would need to be transformed into a comparable unit before they can be combined. The commonly used method of combining outcomes

that measure the same underlying construct with different scales is the standardised mean difference (SMD).

The standardised mean difference standardises the results across trials allowing them to be combined in a meta-analysis. The standardisation is obtained by dividing the difference in mean outcome between two groups with the pooled standard deviation of the measurement. The result of this calculation is that the outcome is measured in standard deviation units. These can be difficult to interpret and a rule of thumb has been suggested, where a SMD of 0.2 standard deviation units is to be considered a small difference between the intervention groups, 0.5 a moderate difference, and 0.8 a large difference (12). A different way to ease interpretation is to re-express the SMD into a familiar scale by multiplying the SMD with a typical among-person standard deviation for a particular scale. This would give the result as a difference in means on this particular scale. There are different approaches to choosing the among-person standard deviation; one solution is to apply the pooled standard deviation of the baseline scores from the trials included in the meta-analysis another solution is to take a standard deviation from an observational study (3).

NEW CHALLENGES WITH SYSTEMATIC REVIEWS

In 2005, the Cochrane review "Interactive Health Communication Applications for people with chronic disease" (13) was retracted because of errors. The review wrongly concluded that an intervention was harmful when in fact it was beneficial. The errors were caused by misinterpretation of the direction of change for several clinical and behavioural outcomes. The misinterpretations were born largely from issues associated with the use of SMDs in meta-analysis. Previous research regarding the use of SMD had focused on whether the SMD would yield similar results as the original metric (14), how to impute missing standard deviations (15) and on identifying the most appropriate estimator of the SMD and the standard deviation (16;17). There was a clear lack of investigation of the reliability of the use of this outcome measure in meta-analysis and this was the motivation for this PhD project.

Given that the goal for systematic reviews has always been to improve the reliability of conclusions, it is imperative that we concentrate on the potential challenges with the method. The overall goal of this thesis is therefore to examine the use of the standardised mean difference in meta-analysis, with the purpose of investigating the challenges that may arise when the method is used in practice and to examine how reliable the method is. This thesis addresses data extraction issues. The statistical issues associated with this method were not investigated (such as different formulae for estimating SMD).

OBJECTIVES

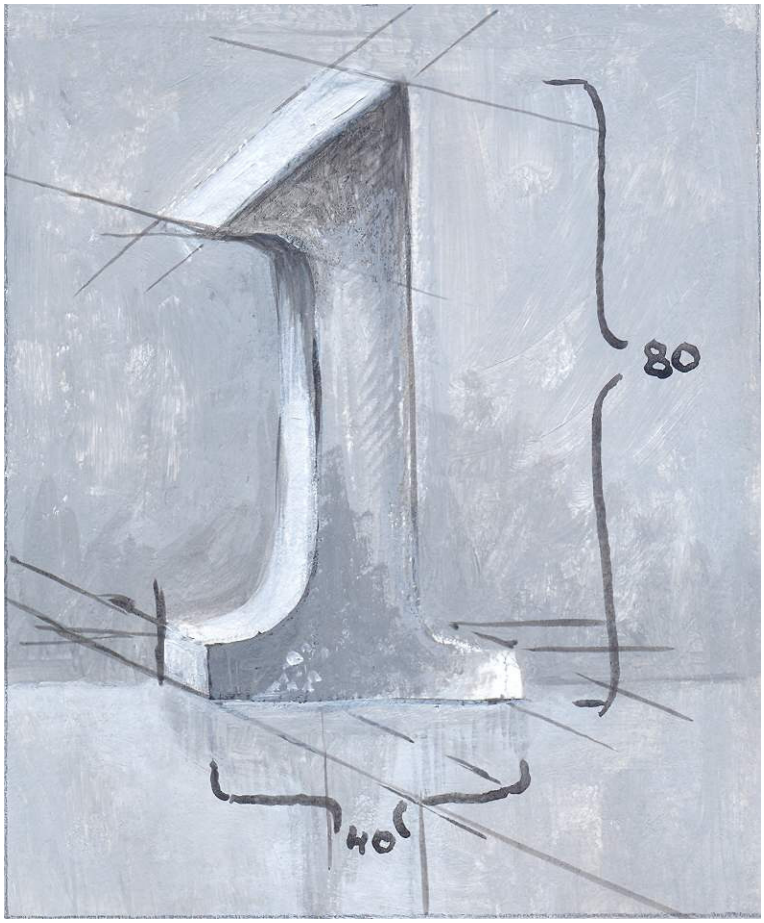
The objectives of this thesis are to investigate the difficulties that may arise when researchers use SMD as an effect measure and to determine the scope for reviewer bias. These issues were explored through three studies addressing the following objectives:

- To investigate whether published meta-analyses using the SMD were accurate (Paper 1).
- To examine the range observer variation when extracting data for the calculation of SMDs (Paper 2).
- To assess the effects of multiple time points, multiple scales of measurement and multiple treatment groups on SMD results (Paper 3).

METHODS & RESULTS

PAPER 1

The aim of this first study was to assess whether SMDs in published meta-analyses were accurate and to describe the nature of the data extraction errors.



Data Extraction Errors in Meta-analyses That Use Standardized Mean Differences

Peter C. Gøtzsche, MD, DrMedSci
Asbjørn Hróbjartsson, MD, PhD
Katja Marić, MSc
Britta Tendal, MSc

Context Meta-analysis of trials that have used different continuous or rating scales to record outcomes of a similar nature requires sophisticated data handling and data transformation to a uniform scale, the standardized mean difference (SMD). It is not known how reliable such meta-analyses are.

Objective To study whether SMDs in meta-analyses are accurate.

Data Sources Systematic review of meta-analyses published in 2004 that reported a result as an SMD, with no language restrictions. Two trials were randomly selected from each meta-analysis. We attempted to replicate the results in each meta-analysis by independently calculating SMD using Hedges adjusted *g*.

Data Extraction Our primary outcome was the proportion of meta-analyses for which our result differed from that of the authors by 0.1 or more, either for the point estimate or for its confidence interval, for at least 1 of the 2 selected trials. We chose 0.1 as cut point because many commonly used treatments have an effect of 0.1 to 0.5, compared with placebo.

Results Of the 27 meta-analyses included in this study, we could not replicate the result for at least 1 of the 2 trials within 0.1 in 10 of the meta-analyses (37%), and in 4 cases, the discrepancy was 0.6 or more for the point estimate. Common problems were erroneous number of patients, means, standard deviations, and sign for the effect estimate. In total, 17 meta-analyses (63%) had errors for at least 1 of the 2 trials examined. For the 10 meta-analyses with errors of at least 0.1, we checked the data from all the trials and conducted our own meta-analysis, using the authors' methods. Seven of these 10 meta-analyses were erroneous (70%); 1 was subsequently retracted, and in 2 a significant difference disappeared or appeared.

Conclusions The high proportion of meta-analyses based on SMDs that show errors indicates that although the statistical process is ostensibly simple, data extraction is particularly liable to errors that can negate or even reverse the findings of the study. This has implications for researchers and implies that all readers, including journal reviewers and policy makers, should approach such meta-analyses with caution.

JAMA. 2007;298(4):430-437 www.jama.com

RESULTS FROM TRIALS THAT HAVE measured the same outcome on the same scale, eg, diastolic blood pressure in mm Hg, can readily be combined in a meta-analysis by calculating the weighted mean difference.[1] Sometimes, trials have used outcomes of a similar nature but that were measured on different scales, eg, pain on a 5-point ranking scale or on a 100-mm visual analog scale, or depression on a clinician-rated scale such as the Hamilton Rating Scale for Depression[2] or a self-rating scale such as the Beck Depression Inventory.[3] In such cases, it is necessary to standardize the measurements on a uniform scale before they can be pooled in a meta-analysis. This is done by calculating the standardized mean difference (SMD) for each trial, which is the difference in means between the 2 groups, divided by the pooled standard deviation of the measurements.[1] By this transformation, the outcome becomes dimensionless and the scales become uniform, eg, for the same degree of pain, values measured on a 100-mm analog scale would be expected to be 20 times larger than values measured on a 5-point ranking scale, but the standard deviation would also be expected to be 20 times larger.

Although simple in principle, it is not known how reliable this method is in practice. In contrast to a meta-analysis of binary data, which usually involves only the extraction of the number of patients and events from the trial reports, a meta-analysis using SMDs requires much more sophisticated data handling, and there are many pitfalls. Standard errors may be mistaken for standard deviations, which will inflate the estimates substantially, and standard deviations may need to be calculated or estimated from *P* values or other data. Some trials may have used changes from baseline instead of values after treatment but may have failed to report data that allow the calculation of within-patient standard deviations. Data extractors also need to know the direction of the scales, which is not always clear in the trial reports. When a high value on one scale means a poor effect, eg, on a depression scale, but a good effect on another scale, eg, a mood scale, it is necessary to change the sign of those values that mean the opposite. Adding to this complexity is that trial

Author Affiliations: Nordic Cochrane Centre, Rigshospitalet, Copenhagen, Denmark.
Corresponding Author: Peter C. Gøtzsche, MD, DrMedSci, Nordic Cochrane Centre, Rigshospitalet, Dept 3343, Blegdamsvej 9, DK-2100 Copenhagen Ø, Denmark (pcg@cochrane.dk).

authors often give changes from baseline as positive values when they should have been negative, eg, when the average value after treatment is lower than the baseline value, or they say they have used changes from baseline when in reality they have used values after treatment. In 1 case, the review authors used the wrong sign for some of the estimates, which led to an erroneous conclusion of harm and retraction of the review, that, when corrected and republished, concluded that the intervention was beneficial.[4]

We studied whether trial SMDs in published meta-analyses are accurate and described the frequency and nature of any data extraction errors and their impact on the meta-analysis result.

METHODS

We performed a PubMed search on March 3, 2005, for meta-analyses that had used the SMD and that were published in 2004. We used the search strategy (*effect size* or *standardised mean difference* or *standardized mean difference* or *SMD*) and (*systematic review* [title and abstract {tiab}] or *meta-analysis* [publication type {pt}] or *review* [pt]). There were no language restrictions.

We included meta-analyses with abstracts that reported an SMD or indicated that there was such a result in the article. The first result in the abstract or in the results section if there was none in the abstract was our index result.

We excluded meta-analyses if (1) the index result was clearly not based exclusively on randomized trials; (2) the index result was based on crossover trials; (3) the index result was not based on at least 2 trials; (4) the authors had used Bayesian statistics; (5) the authors had performed an individual patient data meta-analysis; (6) the meta-analysis had been performed by ourselves; or (7) the meta-analysis was not restricted to humans.

For each meta-analysis, the intervention that appeared to be the authors' primary interest was labeled the experimental intervention. It was easy to determine from the title, introduction, graphs, statistical advice, or grants which intervention was experimental. The other intervention, whether active or inactive, was defined as control. We noted the SMD and its timing for the index result, interventions, disease, any explicit statements about methods for selection of 1 of several possible outcomes or time points in a trial, statistical methods used for pooling, whether values after treatment or changes from baseline had been used, source of funding, and conflicts of interest.

We randomly selected 2 trials from each meta-analysis by using a random numbers table, starting at a new place in the table for every new trial. In one case, the selected trial report could not be retrieved, so we randomly selected another. We extracted outcome data from the trial reports, ensuring that the data extractor on a trial report was different from the one on the corresponding meta-analysis. The trial data extractor was provided with a data sheet with information on the experimental intervention, disease and measurement scale, including any timing if available in the meta-analysis, eg, Hamilton depression score after 6 weeks. Furthermore, the data extractor was informed about the trial result, with its 95% confidence interval (CI), and the group sizes, means and standard deviations for the particular trial's outcome if available, the statistical method used for pooling, and whether final values or changes had been used.

The reason for the lack of blinding was that we wished to see whether we could replicate the published results. We therefore focused on what the authors of the meta-analysis had done and not on what they could have done instead, eg, selected another, perhaps more appropriate, scale when several had been used for measuring the same outcome. Trial data extractors retrieved the necessary information for calculating the SMD from each trial report, including the direction of the effect in relation to the scale used, and could write comments.

Two persons extracted data independently and disagreements (which were mainly caused by simple oversight) were resolved by discussion. We contacted the authors of the meta-analyses for clarification when we could not replicate their data, or when essential data in the trial report for the calculations were missing, ambiguous, or appeared to be erroneous. When the authors had received unpublished data from the trial authors, we used the same unpublished data for our calculations.

Our main outcome was the proportion of meta-analyses for which 1 or both of our 2 trial SMDs differed from that of the authors by 0.1 or more, either for the point estimate or for its CI. We chose 0.1 as the cut point because many commonly used treatments have an effect of 0.1 to 0.5 compared with placebo. For example, the effect of acetaminophen on pain in patients with osteoarthritis is SMD –0.13 (95% CI, –0.22 to –0.04),[5] the effect of antidepressants on mood in trials with active placebos is SMD 0.17 (95% CI, 0.00-0.34),[6] the effect of physical and chemical methods to reduce house dust mite allergens on asthma symptoms is SMD –0.01 (95% CI, –0.10 to 0.13),[7] whereas the effect of inhaled corticosteroids on asthma symptoms is relatively large, SMD –0.49 (95% CI, –0.56 to –0.43).[8] Furthermore, an error of 0.1 can be important when 2 active treatments have been compared, for there is usually little difference between active treatments.

We used Microsoft Excel for our initial calculations of Hedges adjusted g, and *Review Manager*[9] and *Comprehensive Meta Analysis*[10] for our final estimates.

RESULTS

We identified 148 potentially eligible reviews. Fifty-five were excluded based on the abstracts, another 61 after reading the full text, and 5 after reading the 2 randomly selected trial reports (FIGURE 1). The main reasons for exclusion were lack of a reported pooled SMD in the meta-analysis (n=35) or for

DISCUSSION OF PAPER 1

This study gave an impression of the challenges connected with using the SMD. The primary outcome was the proportions of meta-analyses were our result diverged from the authors by more than 0.1 standard deviation units. Looking at two trials per meta-analysis, we found this to be the case for at least one trial in 10 out of 27 meta-analyses. For these ten meta-analyses, we subsequently performed a full meta-analysis including all trials. Seven of the meta-analytic results differed by 0.1 or more. We concluded that meta-analyses using the SMD have a high rate of errors and that could negate or even reverse the findings of the analyses.

This empirical study allowed us to analyse real life results and to get an impression of the practical problems associated with the use of the SMD. A weakness of our method came as a result of the fact that we used a novel approach to determine the reliability of meta-analyses. Because of the extent of the errors we were asked by the editors to do post hoc analyses of the meta-analyses that appeared to be erroneous.

A challenge with the method we used is whether the discrepancies between our estimates and the published estimates could arise from more complex issues than simple errors in data extraction and data handling. Different researchers may hold different views of which data to extract even if the same protocol is followed. This discrepancy becomes an issue if protocols are too open-ended and there are multiple data to choose from in the trial reports. These issues are addressed in the following chapters of this thesis.

Our conclusion is supported by other studies that have examined data extraction, such as a study recently performed by Ford et al. were the authors assessed the conduct of systematic reviews of pharmacological interventions for irritable bowel syndrome (18). They analysed the eligibility criteria and data extraction for dichotomous outcomes in eight systematic reviews and found errors in all eight, leading to errors in 15 out of 16 meta-analytical results. Five of the results differed by more than 10% in the relative treatment effect, when the results were recalculated and in four cases the statistical significance changed.

A study by Horton et al. found error rates around 30%. The study was a cross-sectional study including eighty-seven participants; the participants were amongst other tasks asked to extract six different outcomes from three studies. The authors calculated a reference standard to which the participants' results were compared. Averaged across the six outcomes, 12% of the participants' results differed in statistical significance from the reference standard. These results are in line with our findings.

PAPER 2

In this second study, we focused on the effect of having multiple observers extracting data based on the same protocols and trial reports.

DISCUSSION OF PAPER 2

In this study, we explored the effect of having different observers extracting data. We looked at the impact of the data extraction at two levels. First, we examined how the SMDs for the individual trial results were affected. Second, we examined how the pooled meta-analytical SMDs were affected. We included 10 meta-analyses and 45 trials. To calculate a measure of agreement, we paired the results from the observers in all possible ways (yielding 45 pairs per trial or meta-analysis). For the 45 trials´, 53% of the pairs agreed, meaning that their SMD point estimate and 95% confidence interval differed by less than 0.1. For the 10 meta-analyses, 31% of the pairs agreed. The median size of disagreement was a SMD of 0.22 (Inter quartile range 0.07 to 0.61). The reasons for disagreement were: differences in selection of time points, scales, control groups and type of calculations, whether or not to include a trial in the meta-analysis, and data extraction errors.

We concluded that disagreements were frequent and potentially large. As a consequence, we again recommended that SMD results should be interpreted with caution. We also judged that more detailed protocols would be beneficial, as would statistical expertise and having more than one observer.

The experimental nature of the method we used allowed us to mimic a real life setting and we were able to demonstrate that there were many other problems than simple errors when researchers extract data. Different choices impacted significantly on the results at trial level as well as on meta-analytical results. Not all of the disagreements on trial level were carried over to the meta-analyses, because some of the disagreeing results cancelled each other out. This finding could be interpreted as demonstrating the presence of unsystematic or random error. In our study, none of the researchers had any particular interest in the actual outcomes. It seems unlikely that they would be biased in a particular direction, which supports the theory that error in data extraction has a significant random component. It is important to note that our study also showed that the process of data extraction often involved individual choice. In real life where the observers may have stronger preconceived notions regarding the direction of the effect, decisions may be more prone to selection bias.

It might be seen as a weakness of the study that the observers were only allowed to choose one outcome, although they sometimes found themselves in situations where they wanted to include several outcomes, for example multiple scales (all in accordance with the trial protocols). Had the observers been allowed to include everything they desired (to be on the “safe side”), there might have been more (partial) overlap between pairs. Reporting everything included in the trial reports, however, is far from an optimal way to summarise evidence. The resulting reviews would require complex statistical analyses making the result difficult to interpret in a clinical setting or would be unmanageable and affected by many post hoc decisions since trial reports might include several measures of the same outcome, for example depression scales. Another important issue is that outcomes reported in trial reports might be influenced by selection bias, meaning that they are selected based on statistical significance or magnitude of effect (9;10;19;23). A vital part of performing

a review should be to make decisions in advance on what to extract and summarise. These decisions should include judgements about which outcomes are primary and which methods are considered the best to measure these outcomes. It is sometimes possible to construct hierarchies regarding scales that have been empirically shown to have higher validity than other scales (20).

A potential weakness was that the observers were not allowed to contact the authors. Only reviews in which it was stated that trial authors had not been contacted were included. As the observers started on the data extraction, it became clear that there may have been some unclear reporting of this issue in the reviews, since some of the trial reports had missing data and ambiguities, which made several of the observers express the wish to contact the trial authors. Given our protocol, and the practicalities involved, observers were instructed to exclude a trial if they did not feel comfortable including the trial without contacting the trial authors.

We included both highly experienced observers as well as less experienced observers. We did not draw any conclusion regarding the impact of the level of experience of the observers as our study included too few observers to test any differences. We did find that in cases where calculation was needed, the less experienced observers tended to exclude the studies and the more experienced observers chose to include them and to impute the missing data. This is pointing in the direction that in cases with challenging or missing data, some level of experience is required if data are to be included.

The study by Horton et al. examined the impact of experience on accuracy in data extraction (21). They concluded that experience did not increase accuracy. These authors included three trials, which were selected because all relevant outcomes were reported and there was apparently no need for the observers to perform any calculations. Horton et al.'s conclusion, therefore, may not be applicable to the population of studies that we examined.

PAPER 3

This is the third study. We explored the multiplicity of data that were available for the calculation of an SMD based on a specific outcome.

Title:

Multiplicity of data in trial reports creates an important challenge for the reliability of meta-analyses: an empirical study

(Submitted)

Authors and affiliations:
Britta Tendal (MSc, PhD student)[1], Eveline Nüesch (MSc, PhD, Research Fellow)[2,3] Julian P. T. Higgins (BA, PhD, Senior statistician)[4], Peter Jüni (MD, Head of Division)[2,3], and Peter C. Gøtzsche (MD, DrMedSci, Director)[1]

[1]The Nordic Cochrane Centre, Rigshospitalet, Copenhagen, Denmark
[2]Institute of Social and Preventive Medicine, University of Bern, Switzerland
[3]CTU Bern, Bern University Hospital, Switzerland
[4]MRC Biostatistics Unit, Cambridge, United Kingdom

Corresponding author:

Britta Tendal (MSc, PhD student)

Nordic Cochrane Centre

Rigshospitalet, Dept 3343

Blegdamsvej 9

DK-2100 Copenhagen Ø

Denmark

E-mail: bt@cochrane.dk

Tel: +45 35 45 71 46

Fax: +45 35 45 70 07

Author information

Corresponding Author: Britta Tendal

Author Contributions: All authors had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: Tendal, Nüesch, Gøtzsche

Acquisition of data: Tendal, Nüesch

Analysis and interpretation of data: Higgins, Nüesch, Tendal

Drafting of the manuscript: Tendal, Nüesch

Critical revision of the manuscript for important intellectual content: All authors

Administrative, technical, or material support: Gøtzsche

Study supervision: Gøtzsche

Financial Disclosures: None reported.

Funding/Support: This study is part of a PhD funded by IMK Charitable Fund.

Role of the Sponsors: The funding organizations played no role in the study design and conduct of the study, in the data collection, management, analysis, and interpretation of the data, or in the preparation, review, or approval of the manuscript.

Abstract

Context: Authors performing meta-analyses of clinical trials often face a multiplicity of data in the trial reports. There may be several time points and intervention groups, and the same outcome can be measured on different, but similar scales. The challenge of data multiplicity has not yet been examined in relation to meta-analyses.

Objectives: To examine the scope for multiplicity in a sample of meta-analyses using the standardised mean difference (SMD) as effect measure, and to examine the impact of the multiplicity on the results.

Data source and study selection: We selected all Cochrane reviews published in The Cochrane Library during one year (issues 3, 2006 to 2, 2007) that presented a result as an SMD. We retrieved the trial reports that corresponded to the first SMD result in each review and the review protocols. These index SMDs were used to identify a specific outcome for each meta-analysis from its protocol.
Data Extraction: Based on the protocols and the index outcome, two observers independently extracted the data necessary to calculate SMDs from the trial reports for any intervention group, time point or outcome measure compatible with the protocol. Based on the extracted data, all possible SMDs for the meta-analyses were calculated in Monte Carlo simulations.

Results: Nineteen meta-analyses (83 trials) were included. The review protocols often lacked information about which data to choose. Twenty-four (29%) trials reported data on multiple intervention groups, 30 (36%) provided data on multiple time points and 28 (34%) reported the index outcome measured on multiple scales. In 18 of the 19 meta-analyses, we found multiplicity of data in at least one trial report. In these 18 cases, the median difference between two randomly selected SMDs within the same meta-analysis was 0.11 standard deviation units (range 0.03 to 0.41).

Conclusions: Multiplicity can impact importantly on meta-analyses. To reduce the risk of bias in reviews, protocols should pre-specify which results are preferred in relation to time points, intervention groups and scales.

Introduction

Meta-analyses of randomized clinical trials are pivotal for making evidence-based decisions. There is often multiplicity of data in trial reports regarding multiple intervention groups, multiple time points, multiple outcome measures, and subgroup analyses.[1] This multiplicity is a challenge to meta-analysts, which has received little attention. The choice of the outcome of interest is generally based on clinical judgement. However, a fundamentally similar outcome can be measured on different scales and standardization to a common metric is therefore required before the outcome can be combined in the meta-analysis. This is typically achieved by calculating the standardized mean difference (SMD) for each trial, which is the difference in means between the two groups, divided by the pooled standard deviation of the measurements.[2] By this transformation, the outcome becomes dimensionless and the scales become comparable, as the results are expressed in standard deviation units. For example, a meta-analysis addressing pain might include trials that measured pain on a visual analogue scale and trials that used a 5-point numeric rating scale. This possibility of combining outcomes measured on different scales potentially adds a layer of multiplicity, as the outcome of interest may be measured on more than one scale not only across trials but also within the same trial. Multiplicity of data in trial reports might lead to data driven decisions about what data are included in the meta-analysis and hence is a potential threat to the validity of meta-analysis results.

In this study, we empirically assessed the effect of multiple time points, multiple scales and multiple treatment groups on SMD results in a randomly selected sample of Cochrane reviews.

Methods

Material

We selected all new Cochrane reviews, published in The Cochrane Library during one year (Issues 3, 2006 to 2, 2007) that presented a result as an SMD. We retrieved the reports of all randomised trials that contributed to the first SMD result in each review, and retrieved the latest protocols for all reviews (downloaded in June 2007). Reviews were eligible if the SMD result was based on 2-10 randomized trials and if the outcome was included in the review protocol. Reviews were excluded if only subgroup results were presented. The first pooled SMD result in each review that was not based on a subgroup result was selected as our index SMD result. The index SMD result had to be based on published data only, i.e. there was no indication in the review that the review authors had received additional outcome data from the trial authors. These index SMD results identified a single outcome for each meta-analysis. Following the published protocol, two observers (BT, EN) independently extracted all possible and reasonable data from the trial reports that could be used to calculate the desired SMD for this outcome. For each trial report, we extracted data on all experimental or control groups, time points, and measurement scales, provided they were compatible with the definitions in the protocol. If some required data were unavailable, we used approximations as previously

described.[3] Interim analyses were not included. Disagreements were resolved by discussion. We did not contact trial authors for unpublished data.

Data synthesis

We conducted Monte Carlo simulations for each meta-analysis. To estimate the impact of overall multiplicity, in each trial we randomly sampled one SMD and its corresponding standard error from all possible SMDs generated by all multiple reported data to calculate pooled SMDs using fixed- or random-effects models, as originally done in the published reviews. In each meta-analysis we examined the distribution of pooled SMDs across 10,000 simulations using histograms. To estimate the impact of a single source of multiplicity (intervention groups, time points, measurement scales), we allowed only one source of multiplicity to vary at a time when randomly sampling SMDs for each trial. The other sources of multiplicity were standardized at pre-specified standard values (groups: pooled groups, time point: post treatment, scale: first scale mentioned in text). For example, in the analysis regarding multiplicity originating from scales, the analysis was based on post treatment values and pooled groups (if there were several possible groups). The values of the different scales for this time point and these groups were then randomly sampled for the calculations of the pooled SMD results. The variability of SMD results due to multiplicity across possible variants of a meta-analysis was expressed as the empirical standard deviation of the distributions of pooled SMDs results obtained from the Monte Carlo simulations. Meta-analyses only including trials without multiple data did not contribute to these analyses.

Results

Figure 1 shows the flowchart for the selection of meta-analyses. Of 32 potentially eligible systematic reviews, we excluded 8 because no pooled SMD index result could be selected, 2 because all SMD results were based on unpublished data, 1 because only subgroup results were reported, 1 because no protocol was available and 1 because the SMD result was not described in the protocol. The 19 eligible meta-analyses included 83 trials that contributed to our study.[4-22]

Figure 1. Flowchart for selection of meta-analyses.

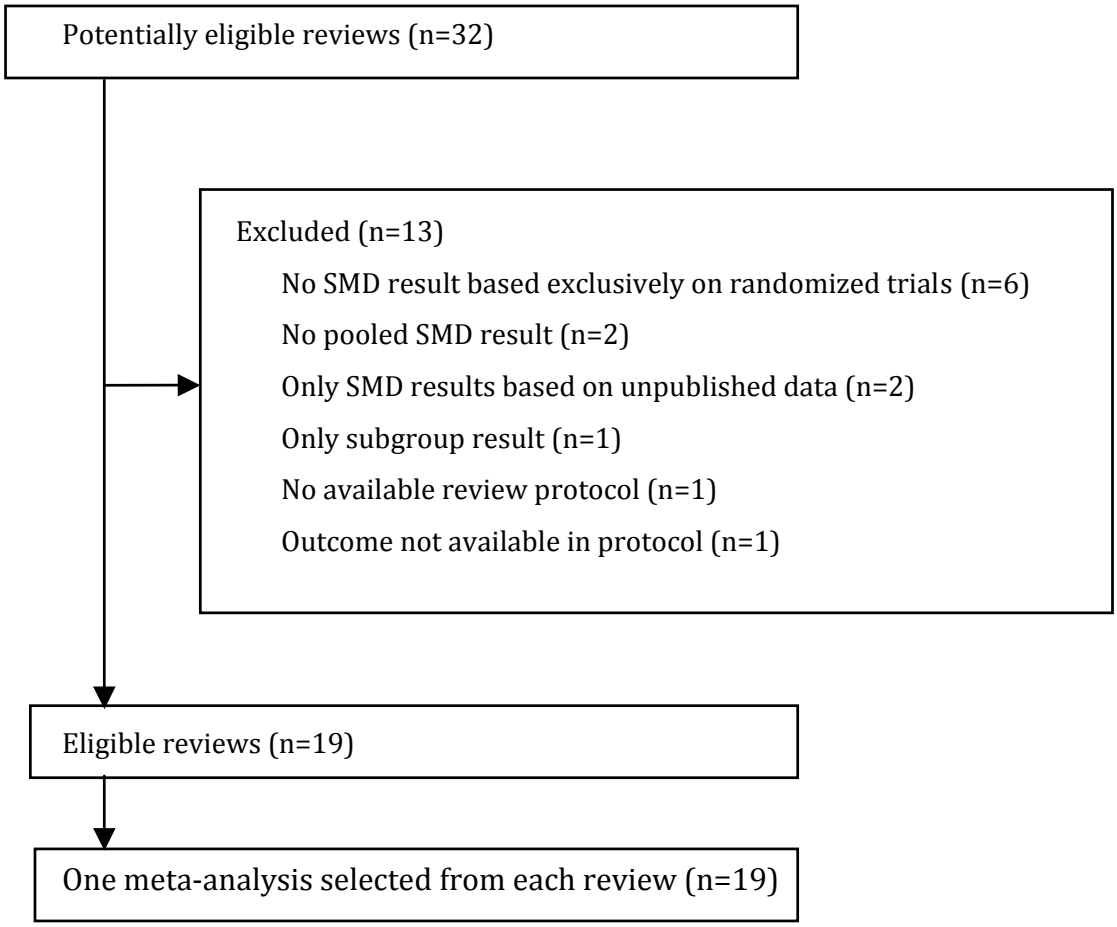


Table 1 shows the characteristics of included reviews. Eight reviews addressed a psychiatric condition, 2 a musculoskeletal condition, 2 a neurological condition, 1 each a gynaecologic, hepatologic and respiratory condition, respectively, and 4 addressed other conditions. Psychological interventions were studied in 10 meta-analyses, pharmacological interventions in 4, physical interventions in 3 and other interventions in 2 meta-analyses (exercise and humidified air). The outcomes analyzed in the 19 meta-analyses were diverse: in 3 meta-analyses the index outcome was pain, in 13 it was another symptom and in 3, other outcomes.

Table 1. Characteristics of included systematic reviews

Author	Outcome	Condition	Intervention	Group
Mytton et al.[16]	School responses	Aggression/violence	Violence prevention program	Cochrane Injuries Group
Afolabi et al.[5]	Neonatal neurological and adaptive score	Caesarean section	Epidural	Cochrane Pregnancy and Childbirth Group
O'Kearney et al.[17]	Depression	Obsessive compulsive disorder	Behavioural therapy or cognitive-behavioural therapy	Cochrane Depression, Anxiety and Neurosis Group
Buckley and Pettit[7]	General functioning score	Schizophrenia	Supportive therapy	Cochrane Schizophrenia Group
Abbass et al.[4]	Anxiety/depression	Common mental disorders	Psychotherapy	Cochrane Depression, Anxiety and Neurosis Group
Orlando et al.[18]	Radiological response	Non-alcoholic fatty liver disease	Bile acids	Cochrane Hepato-Biliary Group
Mistaen and Poot[14]	Patient disease knowledge or symptom management	Postdischarge problem	Telephone follow-up	Cochrane Consumers and Communication Group
Moore and Little[15]	Symptom score	Croup	Humidified air	Cochrane Acute Respiratory Infections Group
Yousefi-Nooraie et al.[22]	Low-back-related disability	Low-back pain	Low level laser therapy	Cochrane Back Group
Trinh et al.[19]	Pain	Neck disorder	Acupuncture	Cochrane Back Group
Martinez Devesa et al.[13]	Subjective tinnitus loudness	Tinitus	Cognitive behavioural therapy	Cochrane Ear, Nose and Throat Disorders Group
Ahmad et al.[16]	Pain	Hysterosalpingography (tubal patency)	Analgesic	Cochrane Menstrual Disorders and Subfertility Group
Woodford and Price[21]	Range of movement	Stroke	EMG biofeedback	Cochrane Stroke Group
Larun et al.[12]	Anxiety	Anxiety	Exercise	Cochrane Depression, Anxiety and Neurosis Group
Gava et al.[9]	Symptom level	Obsessive compulsive disorder	Psychological treatment	Cochrane Depression, Anxiety and Neurosis Group
Furukawa et al.[8]	Global judgement	Panic disorders	Combined treatment: Psychotherapy and antidepressant	Cochrane Depression, Anxiety and Neurosis Group
Ipser et al.[11]	Symptom severity scale	Treatment-resistant anxiety disorders	Pharmacotherapeutic augmentation	Cochrane Depression, Anxiety and Neurosis Group
Uman et al.[20]	Pain	Needle-related procedural pain and distress	Psychological interventions	Cochrane Pain, Palliative and Supportive Care Group
Hunot et al.[10]	Worry/fear symptoms	Generalised anxiety disorder	Psychological therapies	Cochrane Depression, Anxiety and Neurosis Group

Information in the review protocols

The level of information in the review protocols is given in Table 2. None of the review protocols contained information on which scales should be preferred. Eight protocols gave information about which time point or period to select, but only one gave enough information to avoid multiplicity, as the outcome was post treatment. A typical statement leaving much room for data-driven decisions regarding the selection of a time point was: "All outcomes were reported for the short term (up to 12 weeks), medium term (13 to 26 weeks), and long term (more than 26 weeks)".[7] Another example was a review regarding humidified air for treating croup,[15] which stated, "The outcomes will be separately recorded for the week following treatment." The selected outcome was croup symptom score and none of the three included trials ran for so long time but reported symptoms from 20 min to 12 hours after the intervention. Eighteen protocols described which type of control group to select but none reported any hierarchy among similar control groups or any intention to combine such groups.

Table 2. Content of the review protocols.

	Mytton et al.[16]	Afolabi et al.[5]	O'Kearney et al.[17]	Buckley and Pettit[7]	Abbass et al.[4]	Orlando et al.[18]	Mistaen and Poot[14]	Moore and Little[15]	Yousefi-Nooraie et al.[22]	Trinh et al.[19]	Martinez Devesa et al.[13]	Ahmad et al.[16]	Woodford and Price[21]	Larun et al.[12]	Gava et al.[9]	Furukawa et al.[8]	Ipser et al.[11]	Uman et al.[20]	Hunot et al.[10]
Eligible intervention groups	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Eligible control groups	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
Hierarchy of control groups		✓*						✓*											
Eligible time points	✓			✓	✓		✓	✓				✓				✓			✓
Hierarchy of time points																			✓
Eligible measuring methods or scales	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Hierarchy of measuring methods or scales																			

*Only 1 possible control group stated

Observed multiplicity in trial reports

Table 3 presents the extent of multiplicity observed in the 19 reviews including 83 trials. Across all reviews 55 (66%) trials had multiple data from one or more of the three sources. Twenty-four (29%) trials reported data on more than one intervention group or more than one control group, 30 (36%) trials provided data on more than one eligible time point and 28 (34%) trials reported the index outcome using more than one eligible measurement scale.

Table 3 Observed multiplicity in the meta-analyses

		No trials with multiplicity of data regarding:			
	No trials included	**Any of the three source**	**Intervention groups**	**Time points**	**Measurement scales**
Mytton et al.[16]	2	1 (50%)	1 (50%)	0 (0%)	0 (0%)
Afolabi et al.[5]	2	1 (50%)	0 (0%)	1 (50%)	0 (0%)
O'Kearney et al.[17]	2	1 (50%)	1 (50%)	0 (0%)	0 (0%)
Buckley and Pettit[7]	2	2 (100%)	0 (0%)	2 (100%)	0 (0%)
Abbass et al.[4]	2	2 (100%)	0 (0%)	1 (50%)	2 (100%)
Orlando et al.[18]	3	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Mistaen and Poot[14]	3	1 (33%)	1 (33%)	0 (0%)	0 (0%)
Moore and Little[15]	3	2 (67%)	0 (0%)	2 (67%)	0 (0%)
Yousefi-Nooraie et al.[22]	3	2 (67%)	0 (0%)	1 (33%)	1 (33%)
Trinh et al.[19]	3	3 (100%)	0 (0%)	3 (100%)	1 (33%)
Martinez Devesa et al.[13]	4	4 (100%)	3 (75%)	3 (75%)	1 (25%)
Ahmad et al.[16]	5	3 (60%)	1 (20%)	2 (40%)	1 (20%)
Woodford and Price[21]	5	4 (80%)	2 (40%)	2 (40%)	3 (60%)
Larun et al.[12] §	5[§]	4 (80%)	1 (20%)	3 (60%)	2 (40%)
Gava et al.[9]	7	6 (86%)	4 (57%)	3 (43%)	5 (71%)
Furukawa et al.[8]	7	6 (86%)	6 (86%)	2 (29%)	4 (57%)
Ipser et al.[11]	7	6 (86%)	0 (0%)	5 (71%)	3 (43%)
Uman et al.[20]	9	2 (22%)	2 (22%)	0 (0%)	0 (0%)
Hunot et al.[10]	9	5 (56%)	2 (22%)	0 (0%)	5 (56%)
All included reviews	83	55 (66%)	24 (29%)	30 (36%)	28 (34%)

[§]One trial from Larun et al.[12] was excluded because lack of data in the trial report. Meta-analyses are ordered according to the number of trials included.

Observed multiplicity in meta-analyses

In 11 of 19 (58%) meta-analyses, we found at least one trial that provided data on more than one intervention or more than one control group. Thirteen (68%) meta-analyses included at least one trial that reported more than one eligible time point and 11 (58%) meta-analyses at least one trial that reported the index outcome using more than one eligible measurement scale. We found one meta-analysis without multiplicity, as all 3 included trials only reported data of one intervention and control group, one eligible time point and one measurement scale for the index outcome.[18]

Effects of multiplicity on results of meta-analyses

Figure 2 presents distributions of possible pooled SMDs in each meta-analysis, when randomly selecting one possible SMD result per trial. The dots below the distributions indicate how many trials were included in the meta-analyses, open dots are trials without multiplicity, and filled dots are trials with multiplicity. Meta-analyses are ordered according to the number of trials included.
We found that pooled SMD results were affected by any type of multiplicity of data in the included trials in 17 of 19 (89%) meta-analyses, in 1 meta-analysis we did not find multiple data in the trial reports[18] and in 1 meta-analysis the observed multiplicity had no effect on the pooled SMD results.[7] In all 11 (58%) meta-analyses including at least one trial with more than one experimental or control group, we found variability in the pooled SMD results due to this type of multiplicity. In 12 (63%) meta-analyses there was variability in the pooled SMD results due to multiplicity of data regarding time points (Figure 2, 3rd column). In one meta-analysis with two trials that reported more than one eligible time point, we did not find multiplicity due to these different time points.[7] In 9 (47%) meta-analyses we found variability in pooled SMD results from trial data of multiple measurement scales used for the index outcome. In two meta-analyses, one trial in each meta-analysis reported data on more than one measurement scale for the index outcome, but this multiplicity did not affect the pooled SMD results.[6, 22]

Figure 2 Monte Carlo distributions of possible pooled SMDs in each meta-analysis by source of variability. The dots below the distributions indicate how many trials were included in the meta-analyses, open dots are trials without multiplicity of data, and filled dots are trials with multiplicity of data.

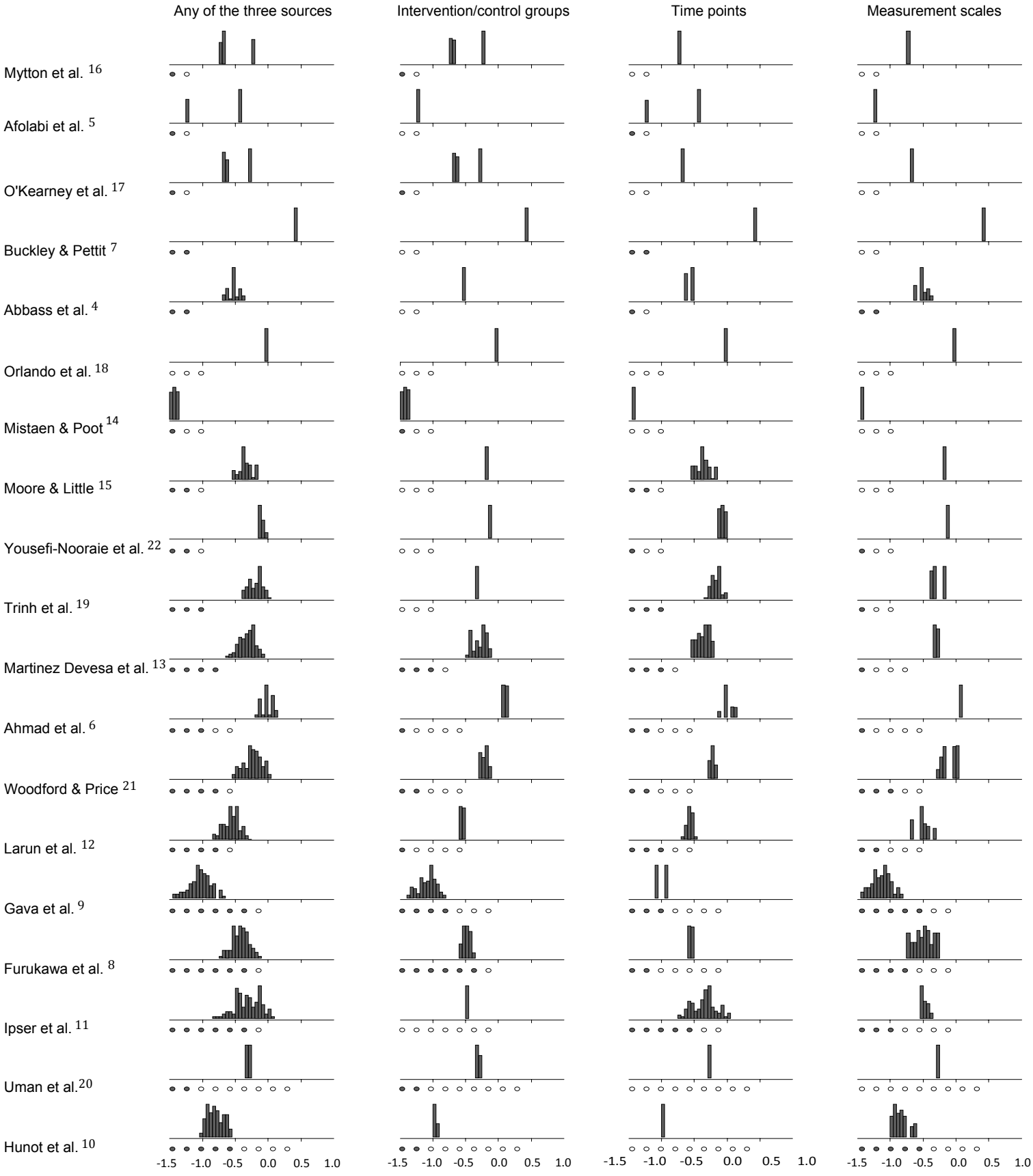


Table 4 presents the variability of pooled SMD results according to different sources of multiplicity. We found 18 meta-analyses that included trials with multiple data for one or more of the three sources evaluated. In these 18 cases, the median difference between two randomly selected SMDs within the same meta-analysis was 0.11 standard deviation units (range 0.03 to 0.41).

Table 4: Variability in meta-analyses results

Source of multiplicity	Number of meta-analyses with multiplicity of data	Variability in SMD results across meta-analyses (standard deviation [range])
Intervention groups	11 of 19 (58%)	0.05 (0.01 to 0.23)
Time points	13 of 19 (68%)	0.06 (0.02 to 0.41)
Measurement scales	11 of 19 (58%)	0.09 (0.01 to 0.15)
Any source	18 of 19 (95%)	0.11 (0.03 to 0.41)

The median difference across the 11 meta-analyses that included trials with multiple data regarding intervention groups was 0.05 standard deviation units (range 0.01 to 0.23) between two randomly selected SMDs calculated from the eligible intervention groups. The median standard deviation across 13 meta-analyses that included trials with data on multiple eligible time points was 0.06 (range 0.02 to 0.41), and across 11 meta-analyses including trials that provided data of multiple measurement scales for the index outcome the median was 0.09 (range 0.01 to 0.15).

Comment

In 18 out of the 19 meta-analyses included in our study, we found multiplicity of data in trial reports in at least one trial, which frequently resulted in substantial variation in the pooled SMD results. The impact of multiple data in trial reports regarding intervention groups, time points or measurement scales on meta-analysis results varied across meta-analyses ranging from essentially no impact to a large one (0.41 standard deviation units), with a median difference of 0.11 standard deviation units. We also estimated the impact of the individual sources of multiplicity, holding the other sources constant.

We randomly selected Cochrane reviews and included a broad selection of interventions and outcomes. The variability of the SMD results did not seem related to particular types of interventions or outcomes. To estimate the impact of multiplicity on meta-analysis results, we randomly selected one SMD per trial from a pool of eligible SMDs with equal probability. This approach explores what is possible due to multiple data. However, there might be implicit rules regarding data extraction within specialties. For example, one scale might be more commonly used, e.g. Hamilton's depression scale, than other scales. Such implicit hierarchy of scales would be expected to reduce the multiplicity, but should be made explicit in protocols for systematic reviews.

Our results are transparent as we only included published results. This means that we likely underestimated the true level of multiplicity, as selective reporting of outcomes in trials is common.[23-26] Positive, statistically significant results are more likely to be published than non-significant results.[27]

Our study was possible because authors of Cochrane Reviews are required to publish their protocols before they embark on the review. We believe that for most meta-analyses published outside the Cochrane Library, no protocol is available,[28] and the scope for multiplicity is therefore likely greater. We examined three frequent sources of multiplicity of data in trial reports: intervention groups, time points, and measurement scales. There are other types of multiple data in trial reports. For example, results might be reported for different types of analyses: intention-to-treat and per-protocol analyses. For each meta-analysis, there can also be several other outcomes than our index outcome, which might be selected according to whether or not the result appears favourable.

The extent of multiplicity of data found in trial reports is a function of the information provided in the review protocols: a poorly specified outcome will be expected to lead to more multiplicity. Some might argue that data extraction for a meta-analysis is dependent on what is reported in trials and cannot be entirely specified in advance without knowledge of the included trials. However, we argue that to minimise data-driven selection of time points, measurement scales or intervention groups, researchers should specify these decisions at the protocol stage. If amendments to the protocol are indicated, these should be transparently reported.[29, 30]

To our knowledge, our study is the first to show empirically the extent to which the reliability of meta-analysis results may be compromised by multiplicity of data. We have previously reported results from an observer agreement study performed on 10 of the meta-analyses included in this study.[2] We found that disagreements among observers were common and often large, the main reasons for disagreement being different choices of groups, time points, scales and calculations, different decisions on in- or exclusion of certain trials and data extraction errors.[2] A recent paper by Bender et al. describes the problem of multiple comparisons in systematic reviews.[1] The authors identified common reasons for multiplicity in reviews, but did not estimate the impact on the meta-analytic results.[1] In our study, we included meta-analyses of SMDs, which may be particularly associated with multiplicity of data due to the use of different measurement scales in included trials. However, multiplicity of data due to selection of time points and groups is not unique to SMD, and future studies could therefore explore whether multiplicity is also an issue for other effect measures, including binary outcomes.

One approach towards dealing with multiplicity in systematic reviews is to extract, analyse and report all data available on intervention groups, time points and measurement scales. However, this may lead to considerable problems with interpretation in view of potential discrepancies between different scales or different time points. In analogy to repetitive measures in an individual trial, all available time points reported in included trials could be analysed in a single meta-analysis while fully accounting for the correlation of repetitive measurements within a trial.[31] Or, like the use of bivariate

models used in diagnostic research,[32] assessments from different scales measuring similar concepts could be analysed in a single multivariate model. Whereas the first approach of including repetitive assessments in a single analysis may be easily understandable, the second approach will appear cryptic to many readers.

A better approach seems to be to address the issues of multiplicity by providing more detailed protocols for systematic reviews, with clearly specified time points, scales and groups to consider and explicit and transparent hierarchies to be used in case of multiplicity of scales, groups or time points. Clinical judgment will be important here. Ideally, the choice of time points and scales should be evidence based, but empirical evidence for the most interesting time points and a hierarchy of scales according to their validity and responsiveness are rarely available. In addition, it is difficult to foresee everything at the protocol stage and the scope, the methodological quality, and the quality of reporting of included studies might require subsequent modifications.[33] Only Cochrane reviews are formally required to have a published protocol, however, and only around ten percent of non-Cochrane reviews explicitly stated to be based on a formal protocol.[30] Protocol amendments may impact on results and conclusions of systematic reviews and should be made only after careful consideration and be reported transparently.[29, 30]

Conclusions

Variability in meta-analysis results related to the multiplicity of data in trial reports and to review protocols lacking a detailed specification of eligible time points, scales and treatment groups is substantial. Systematic reviews are studies in their own right and reviewers should anticipate multiplicity of data in trial reports and take this into account when writing protocols. To enhance reliability of meta-analyses, we suggest that protocols should clearly define time points to be extracted, give a hierarchy of scales, clearly define eligible treatment and control groups, and give strategies for handling multiplicity of data.

References

1. Bender R, Bunce C, Clarke M, et al. Attention should be given to multiplicity issues in systematic reviews. *J Clin Epidemiol.* Sep 2008;61(9):857-865.

2. Tendal B, Higgins JP, Juni P, et al. Disagreements in meta-analyses using outcomes measured on continuous or rating scales: observer agreement study. *BMJ.* 2009;339:b3128.

3. Reichenbach S, Sterchi R, Scherer M, et al. Meta-analysis: chondroitin for osteoarthritis of the knee or hip. *Ann Intern Med.* Apr 17 2007;146(8):580-590.

4. Abbass AA, Hancock JT, Henderson J, Kisely S. Short-term psychodynamic psychotherapies for common mental disorders. *Cochrane Database Syst Rev.* 2006(4):CD004687.

5. Afolabi BB, Lesi FE, Merah NA. Regional versus general anaesthesia for caesarean section. *Cochrane Database Syst Rev.* 2006(4):CD004350.

6. Ahmad G, Duffy J, Watson AJ. Pain relief in hysterosalpingography. *Cochrane Database Syst Rev.* 2007(2):CD006106.

7. Buckley LA, Pettit T, Adams CE. Supportive therapy for schizophrenia. *Cochrane Database Syst Rev.* 2007(3):CD004716.

8. Furukawa TA, Watanabe N, Churchill R. Combined psychotherapy plus antidepressants for panic disorder with or without agoraphobia. *Cochrane Database Syst Rev.* 2007(1):CD004364.

9. Gava I, Barbui C, Aguglia E, et al. Psychological treatments versus treatment as usual for obsessive compulsive disorder (OCD). *Cochrane Database Syst Rev.* 2007(2):CD005333.

10. Hunot V, Churchill R, Silva de Lima M, Teixeira V. Psychological therapies for generalised anxiety disorder. *Cochrane Database Syst Rev.* 2007(1):CD001848.

11. Ipser JC, Carey P, Dhansay Y, Fakier N, Seedat S, Stein DJ. Pharmacotherapy augmentation strategies in treatment-resistant anxiety disorders. *Cochrane Database Syst Rev.* 2006(4):CD005473.

12. Larun L, Nordheim LV, Ekeland E, Hagen KB, Heian F. Exercise in prevention and treatment of anxiety and depression among children and young people. *Cochrane Database Syst Rev.* 2006;3:CD004691.

13. Martinez Devesa P, Waddell A, Perera R, Theodoulou M. Cognitive behavioural therapy for tinnitus. *Cochrane Database Syst Rev.* 2007(1):CD005233.

14. Mistiaen P, Poot E. Telephone follow-up, initiated by a hospital-based health professional, for postdischarge problems in patients discharged from hospital to home. *Cochrane Database Syst Rev.* 2006(4):CD004510.

15. Moore M, Little P. Humidified air inhalation for treating croup. *Cochrane Database Syst Rev.* 2006;3:CD002870.

16. Mytton J, DiGuiseppi C, Gough D, Taylor R, Logan S. School-based secondary prevention programmes for preventing violence. *Cochrane Database Syst Rev.* 2006;3:CD004606.

17. O'Kearney RT, Anstey KJ, von Sanden C. Behavioural and cognitive behavioural therapy for obsessive compulsive disorder in children and adolescents. *Cochrane Database Syst Rev.* 2006(4):CD004856.

18. Orlando R, Azzalini L, Orando S, Lirussi F. Bile acids for non-alcoholic fatty liver disease and/or steatohepatitis. *Cochrane Database Syst Rev.* 2007(1):CD005160.
19. Trinh KV, Graham N, Gross AR, et al. Acupuncture for neck disorders. *Cochrane Database Syst Rev.* 2006;3:CD004870.
20. Uman LS, Chambers CT, McGrath PJ, Kisely S. Psychological interventions for needle-related procedural pain and distress in children and adolescents. *Cochrane Database Syst Rev.* 2006(4):CD005179.
21. Woodford H, Price C. EMG biofeedback for the recovery of motor function after stroke. *Cochrane Database Syst Rev.* 2007(2):CD004585.
22. Yousefi-Nooraie R, Schonstein E, Heidari K, et al. Low level laser therapy for nonspecific low-back pain. *Cochrane Database Syst Rev.* 2007(2):CD005107.
23. Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *BMJ.* Apr 2 2005;330(7494):753.
24. Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA.* May 26 2004;291(20):2457-2465.
25. Chan AW, Krleza-Jeric K, Schmid I, Altman DG. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *CMAJ.* Sep 28 2004;171(7):735-740.
26. Vedula SS, Bero L, Scherer RW, Dickersin K. Outcome reporting in industry-sponsored trials of gabapentin for off-label use. *N Engl J Med.* Nov 12 2009;361(20):1963-1971.
27. Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database Syst Rev.* 2009(1):MR000006.
28. Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. *PLoS Med.* Mar 27 2007;4(3):e78.
29. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ.* 2009;339:b2700.
30. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ.* 2009;339:b2535.
31. Wandel S, Juni P, Tendal B, et al. Glucosamine, chondroitin or placebo for osteoarthritis of the hip or knee: network meta-analysis. 2010.
32. Harbord RM, Whiting P, Sterne JA, et al. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. *J Clin Epidemiol.* Nov 2008;61(11):1095-1103.
33. Juni P, Egger M. PRISMAtic reporting of systematic reviews and meta-analyses. *Lancet.* Oct 10 2009;374(9697):1221-1223.

DISCUSSION OF PAPER 3

In this study, we examined the extent of multiplicity of data that could be used to calculate an SMD based on a specific outcome. Furthermore, we assessed the impact of this multiplicity on the meta-analytical results. Nineteen meta-analyses (83 trials) were included. Twenty-four (29%) trials reported data on multiple intervention groups, 30 (36%) provided data on multiple time points of data assessment and 28 (34%) reported the index outcome measured on multiple scales. In 18 of the 19 meta-analyses, we found multiplicity of data in at least one trial report. In these 18 cases, the median difference between two randomly selected SMDs within the same meta-analysis was 0.11 standard deviation units (range 0.03 to 0.41).

Our conclusion was that there is substantial multiplicity and that it can impact importantly on meta-analyses. Reviewers should anticipate multiplicity of data in trial reports and take this into account when writing protocols. To enhance reliability of meta-analyses, we suggest that protocols should clearly define time points to be extracted, give a hierarchy of scales, clearly define eligible treatment and control groups, and give strategies for handling multiplicity of data.

The method we used allowed us to fully explore the multiplicity of data presented in the trial reports, and our selection of data was based on the inclusion and exclusion criteria defined in the review protocols. In the Monte Carlo simulation, we randomly sampled from the relevant outcomes presented in each trial report. This gave all the data the same chance of being sampled for the pooled SMDs. In real life, the process of selecting data is probably not so random. The process of performing a review is iterative; as the review process advances, the researchers might apply post hoc decision rules that fail to be amended in the protocol for the review. There may also be implicit decision rules within the review group. For example, one review group may always prefer a specific scale or a specific way of handling multiple intervention groups. These choices should be made transparent by adding this information to the review protocols.

We used pre-defined standards (post treatment, pooled groups and first scale mentioned in text) in the Monte Carlo analyses allowing us to analyse the sources of multiplicity individually. Had we chosen different standards, we might have gotten different results. We opted for this solution, as it ensured that we would not let the data influence our choice of standards.

Our recommendations that protocols should be more detailed are theoretical. The purpose of our study was to explore the levels of multiplicity and the effect on the meta-analytical results; it was not directly aimed at finding solutions for multiplicity. I suggest that future research should explore ways to handle this challenge.

In a study by Bender et al. aiming to investigate multiple comparisons within reviews, several sources of multiplicity were identified amongst others multiple time points, groups and outcomes (22).

Bender et al. suggested that multivariate analyses and a priori decisions regarding the preferred outcomes, comparison groups and time points could offer a solution to the multiplicity challenge. They stressed that these were only suggestions and that more research was needed.

In our study, we only focused on one outcome per review. We investigated the variety of combined results that could be obtained by applying the restrictions defined in the review protocols on the included trial reports. This, however, leaves the question open as to how the authors of the reviews handled the multiplicity. What did they choose to report and why? This question could be answered by a study examining the published reviews combined with a survey among the authors of the reviews.

CONCLUSIONS

My overall conclusion is that the SMD is a useful method but that there are challenges, which should be acknowledged and taken into account when one applies this method. In the two first studies, we found that errors occur easily when using SMD as an effect measure and we also demonstrated the ever-present challenge with observer variation. In research, we aim to minimise the effect of observer variation, which can only be achieved after acknowledgement of its presence. In the first two studies, we examined the risk of selective outcome reporting, as a result of observer variation, and in the context of data extraction for meta-analysis, we confirmed that it can often lead to biased results. This risk was substantiated by our third study, where we show that trials generally contain multiple applicable data that potentially yield different results.

How to handle these challenges? When we submitted our paper on observer variation, a reviewer commented that we addressed a real problem and suggested that the obvious solution was to report everything. Based on the results presented in this thesis, it seems clear that analysing and reporting everything, for all scales, groups and time points is not the solution. Such reporting would be impractical and the results could be hard to interpret, as the amount of multiple data in the trial reports may be large. Another problem is that the selective reporting in the trials might be considerable. It might therefore be necessary to focus on how, and by how much selective reporting could have biased the meta-analyses (3;8-10;23).

In the words of Sir Muir Gray, what we need is: "Clean clear knowledge for decision-making". In my point of view, the best solution lies in the clarity and detail of the protocol of a systematic review, defining which scales are preferable (e.g. least prone to bias), which time points and groups are most relevant and how any combining of groups should be done if relevant. Another focus should be more careful data extraction since the risk of error appears to be high.

FUTURE RESEARCH

This thesis has elucidated some of the many challenges connected with the use of the SMD. I see two important directions for future research: the extent of the challenges beyond SMD and the solutions to the challenges.

As to the extent of the challenges beyond SMD, I believe that many of the issues discussed here apply more broadly than just systematic reviews applying the SMD method. The majority of the challenges are connected with data extraction issues that apply generally to systematic reviews. It would therefore be of value to examine these issues in reviews applying different meta-analytical approaches such as the weighted mean difference, odds ratio and relative risk. Also, review authors' methods of handling multiplicity could be investigated by examining published reviews and by performing a

survey among the review authors. The aim would be to investigate how multiplicity is handled and the reasons behind the choices.

The second direction for future research should be to find suitable solutions. As mentioned earlier, I believe that improving the quality of protocols for reviews is an important step. But to draw any firm conclusions it would have to be investigated whether more detailed protocols lead to more reliable systematic reviews. It also requires more research to decide which details belong in a protocol for a systematic review. There are different guidelines for writing protocols for systematic reviews available today (3;5;6). These guidelines differ somewhat in their recommendations for the development and reporting of protocols.

If part of the problem is poor reporting of the protocols, a solution would be to develop an evidence-based reporting guideline aiming at giving recommendations on which information items to include and report on in a review protocol. Such a project has already been initiated for protocols for randomised trials (24). Having a reporting guideline would help to ensure clarity and transparency allowing the reader to better judge the reliability and validity of the study. The guideline should include a checklist, as this would help the authors when they write the protocol, and it would also help the readers identify any missing information in the protocol. Developing a good evidence based reporting guideline would not be an easy task. It would be necessary to involve various stakeholders from the beginning in consensus processes and it would require regular updates of the guideline (25).

REFERENCE LIST

(1) Deeks JJ. Systematic Reviews Evaluating Effects of Health Care Interventions: Issues of Synthesis and Bias. Birmingham, 2007.

(2) D'Agostino RB, Kwan H. Measuring effectiveness. What to expect without a randomized control group. Med Care 1995 (4 Suppl):AS95-105.

(3) Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.2 [Updated September 2009] The Cochrane Collaboration, 2008. www.cochrane-handbook.org

(4) Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. PLoS Med 2007;4(3):e78.

(5) Centre for Reviews and Dissemination. Systematic reviews: CRD's guidance for undertaking reviews in health care. York, University of York, 2009

(6) Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. Systematic reviews of trials and other studies. Health Technology Assessment 1998;2(19).

(7) Scherer RW, Langenberg P, von EE. Full publication of results initially presented in abstracts. Cochrane Database Syst Rev 2007;(2):MR000005.

(8) Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database Syst Rev 2009;(1):MR000006.

(9) Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 2004;291(20):2457-65.

(10) Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. BMJ 2005;330(7494):753.

(11) Williamson PR, Gamble C, Altman DG, Hutton JL. Outcome selection bias in meta-analysis. Stat Methods Med Res 2005;14(5):515-24.

(12) Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale, NJ, Lawrence Erlbaum, 1988.

(13) Murray E, Burns J, See TS, Lai R, Nazareth I. Interactive Health Communication Applications for people with chronic disease. Cochrane Database Syst Rev 2005;(4):CD004274.

(14) Kelley GA, Tran ZV. Original metric versus standardized effect sizes for meta-analysis of clinical data. Prev Cardiol 2001;4(1):40-45.

(15) Furukawa TA, Barbui C, Cipriani A, Brambilla P, Watanabe N. Imputing missing standard deviations in meta-analyses can provide accurate results. J Clin Epidemiol 2006;59(1):7-10.

(16) Cannella KAS. Another decision to evaluate: choice of standardized mean difference effect size estimator [PA06]. 6th Cochrane Colloquium, Baltimore, 22-26 October1998.

(17) White IR, Thomas J. Standardized mean differences in individually-randomized and cluster-randomized trials, with applications to meta-analysis. Clin Trials 2005;2(2):141-51.

(18) Ford AC, Guyatt GH, Talley NJ, Moayyedi P. Errors in the conduct of systematic reviews of pharmacological interventions for irritable bowel syndrome. Am J Gastroenterol 2010;105(2):280-8.

(19) Furukawa TA, Watanabe N, Omori IM, Montori VM, Guyatt GH. Association between unreported outcomes and effect size estimates in Cochrane meta-analyses. JAMA 2007;297(5):468-70.

(20) Juni P, Reichenbach S, Dieppe P. Osteoarthritis: rational approach to treating the individual. Best Pract Res Clin Rheumatol 2006;20(4):721-40.

(21) Horton J, Vandermeer B, Hartling L, Tjosvold L, Klassen TP, Buscemi N. Systematic review data extraction: cross-sectional study showed that experience did not increase accuracy. J Clin Epidemiol 2010;63(3):289-98.

(22) Bender R, Bunce C, Clarke M, Gates S, Lange S, Pace NL, et al. Attention should be given to multiplicity issues in systematic reviews. J Clin Epidemiol 2008;61(9):857-65.

(23) Chan AW, Krleza-Jeric K, Schmid I, Altman DG. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. CMAJ 2004;171(7):735-40.

(24) Chan A-W, Tetzlaff J, Altman DG, Gotzsche PC, Hrobjartsson A, et.al. The SPIRIT initiative: Defining Standard Protocol Items for Randomized Trials. German J Evid Quality Health Care (suppl) 102[s27]. 2008.

(25) Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. PLoS Med 2010;7(2):e1000217.