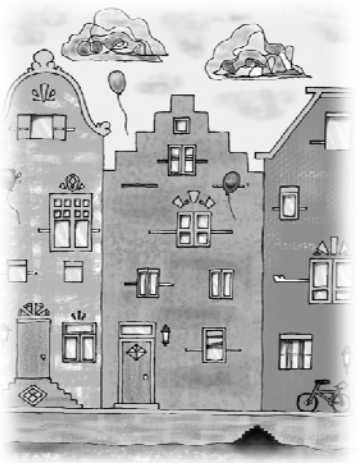
CAUSAL META-ANALYSIS

METHODOLOGY AND APPLICATIONS



Leon Bax

To the memory of
my mother

CAUSAL META-ANALYSIS
METHODOLOGY AND APPLICATIONS.

Utrecht, Utrecht University, Faculty of Medicine
Ph.D. dissertation, with a summary in Dutch
Proefschrift, met een samenvatting in het Nederlands

Printed by Gildeprint
Cover design by Shizuka Tsuruta

I would like to acknowledge the financial contribution of the Julius Center for Health Sciences and Primary Care for the publication of this dissertation.

CAUSAL META-ANALYSIS

METHODOLOGY AND APPLICATIONS

CAUSALE META-ANALYSE

METHODOLOGIE EN TOEPASSINGEN

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. J.C. Stoof, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op dinsdag 7 juli 2009 des middags te 4.15 uur

door

Leendert Joost Bax

geboren op 19 april 1973, te Gorinchem

Promotor: Prof. dr. K.G.M. Moons

Science is built of facts
the way a house is built of bricks...

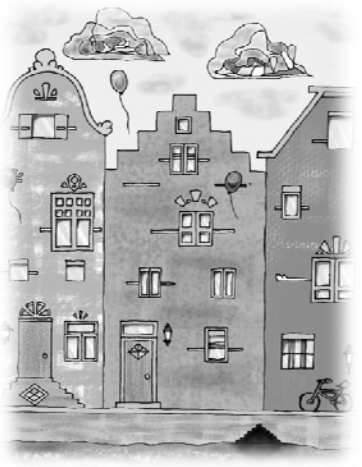
…and an accumulation of facts
is no more science than a pile of bricks is a house.

Jules-Henri Poincaré

French mathematician & physicist (1854 - 1912)

CONTENTS

INTRODUCTION



1

An Introduction to Meta-Analysis

1.1 What is meta-analysis?

The term meta-analysis is relatively new and was first introduced by Gene Glass[1] in 1976. The words meta and analysis are much older and have their etymological origins in Ancient Greek. Meta comes from μετα which meant 'after' or 'beyond', and analysis comes from the verb αναλυω which translates loosely as 'to unravel' or 'take apart'. Following these original words literally, one could deduce that meta-analysis must be some sort of investigation that comes after other investigations, and this is indeed the case: meta-analyses combine numerical data from two or more separate studies in a new aggregate-level framework. Although the formal application of meta-analysis to scientific evidence is a recent trend, the mathematical ideas for combining data from separate studies were already described in various contexts in the early 20th century by Pearson[2], Birge[3], and later Cochran[4].

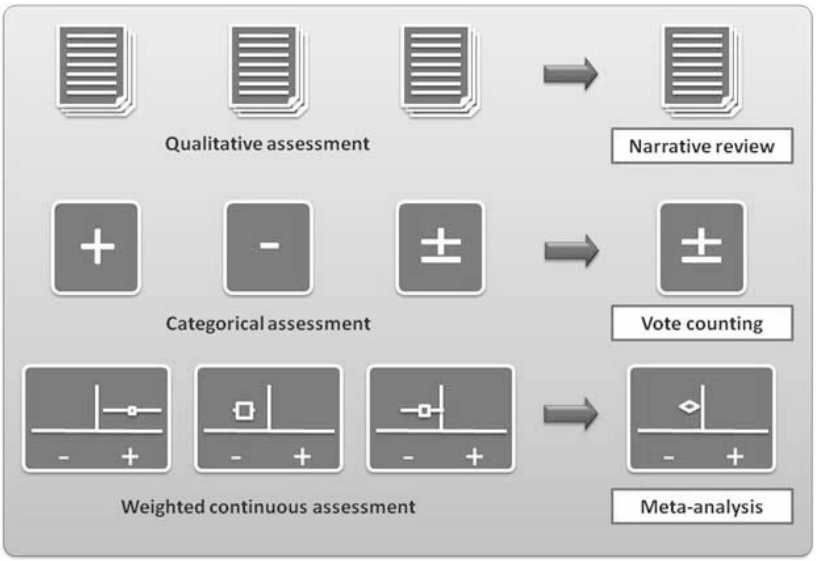


Figure 1.1 From narrative reviews to meta-analysis

Summaries of prior research have evolved from qualitative narrative approaches to sign-based categorical quantifications referred to as vote-counting, and eventually to weighted meta-analysis. The size of the squares in the meta-analysis plots is proportional to the importance (weight) of each study in the analysis. The horizontal bars represent the confidence intervals. The bottom right plot summarizes the evidence of the three studies in the diamond, with the center indicating the summary point estimate and the edges indicating its confidence limits.

Before meta-analysis was popularized in the late 1990s, numerical data from prior studies were generally not combined in a systematic quantitative manner. Instead, they were summed up in a narrative review or combined in a semi-quantitative

approach called vote-counting[5]. In vote-counting, one counts the number of positive, negative, and neutral studies to determine which are most abundant. Unfortunately, this practice disregards the magnitude or strength of the evidence in each study[5], has very low power (which may even decrease as more studies are added[6]), and can lead to misleading results and inferences[6,7]. Proper meta-analysis avoids these problems and provides a summary of numerical results. Figure 1.1 shows the evolution from narrative reviews to vote-counting to meta-analysis.

Although meta-analysis is sometimes used as a synonym for systematic review, most experts in the field reserve the term systematic review for the systematic search, retrieval, and assessment of research reports, and the term meta-analysis for the subsequent quantitative analytical part of the systematic review[8-10]. Hence, it is possible to have a systematic review without a meta-analysis (containing a qualitative instead of quantitative summary of numerical data). It is, however, not recommended to do a meta-analysis that is not incorporated in a systematic review. Applying a meta-analysis in the framework of a systematic review ensures that potential biases can be assessed systematically and accounted for, if possible, in the meta-analysis.

Because data from different sources are combined, meta-analyses have also been referred to as 'pooled analyses' or 'data syntheses'. Interestingly, the latter term also comes from Ancient Greek (συνϑεσις), meaning 'integration' or 'combination', and is in fact the antonym of analysis. This contrast reveals the two major objectives of a proper meta-analysis: (1) to simplify the interpretation of a multitude of data from separate studies by making a numerical synthesis, and (2) to clarify the potential causes of inevitable differences between the studies by differential (stratified or regression) analysis.

We distinguish three types of meta-analysis: those that only deal with summary data from primary studies *j*, those that analyze individual-level data from primary studies *i*, and those that include both types of data. The first type is a classical meta-analysis; the second is often referred to as an individual patient/participant data (IPD) meta-analysis. If both types of data are used within a single (hierarchical) framework, it is referred to as a combined data meta-analysis. The IPD and combined approaches give the researcher more opportunities for investigation. However, they are still relatively uncommon because individual data are not always obtainable and because these projects are time-consuming and complex in their analysis.

1.2 Meta-analysis in medicine: for which gnosis?

Meta-analysis is most predominantly used in the medical and social sciences. As such, it can describe primary studies over the entire spectrum of medical research. A common method to categorize these studies is based on the type of knowledge (gnosis) the study's evidence is supposed to contribute. Such knowledge can be diagnosis, etiognosis, and prognosis[11]. Within this knowledge-based classification, the studies can be categorized according to the type of association that is being investigated: causal associations and descriptive associations[12].

Diagnosis is an ad-hoc concept that refers to knowledge on the probability of having or not having an illness. This probability is naturally based on a number of factors or tests, which are assumed to be associated with the presence or absence of the disease. With illness status as the unknown quantity, this association is clearly a descriptive one and tests are not causal agents in the onset of the illness at hand.

Etiognosis is knowledge on why an illness has occurred, which is a causal concept per se. Etiognostic research investigates whether a certain factor or factors may have illness as a consequence. Where diagnosis is typically best investigated with a cross-sectional approach, etiognostic research requires a longitudinal approach that involves a passage of time. Since the causality in etiognosis is pathogenic, it is mostly unethical to perform etiologic experiments on causal risk-increasing factors. Hence, etiognostic or etiologic research is almost always observational.

Prognosis deals with knowledge on the future course of a health status, and a distinction is made between intervention prognosis and descriptive prognosis[12,13]. Intervention-prognostic research typically investigates causal associations between a therapeutic agent and a health status outcome. In contrast with etiognostic research, however, the causality is not pathogenic but rather physiogenic (expected to improve health). This means experiments that assign people to certain interventions may be possible and ethical, and the ultimate study design for this purpose is the randomized trial. Descriptive prognostic research is not concerned with causation but rather with prediction. The question is not whether certain factors are causing a certain outcome but whether knowledge of these factors can improve knowledge of an individual's future health status.

A meta-analysis can involve all kinds of gnostic research. Most meta-analyses, however, deal with causal (intervention-prognostic and etiognostic) research, although systematic reviews of univariable diagnostic and descriptive prognostic

studies are becoming more and more common. Synthesis of multivariable diagnostic and prognostic studies is an area that is still largely in development.

1.3 Frameworks for statistical inference

Meta-analysis involves statistical inference, meaning that imperfect or only part of all potential data are used to infer something about a much larger construct. This implies uncertainty. It can be argued that an important aim of any analysis is to quantify this uncertainty. Measures of observed data are often referred to as statistics and their corresponding unknown counterparts in a population (that is beyond our investigation) are called parameters. The uncertainty we have about how well our statistics estimate the parameters is quantified with probabilities. There are two major views of probability; one is referred to as Frequentist and the other as Bayesian.

In the Frequentist framework, scientific probability is formally defined as the frequency with which the outcome of interest in a study would occur in a hypothetical sample of such studies. Frequentist analyses often proceed by the definition of a null-hypothesis and an attempt to refute it in Popperian style by assessing whether the probability that the study's results could have occurred if the hypothesis were true is small enough to conclude that it may be based on chance (with 0.05 as the generally recommended cut-off point for this probability). Due to the limited usability and informativeness of such probability statements (formally known as *P* values), Frequentist analyses have evolved more and more to estimation via confidence intervals. The ranges of these Frequentist confidence intervals are, however, not directly inferring to the underlying parameters, but rather indicate a possible parameter range that is compatible with the study's results and would be observed in a certain percentage (e.g. 95%) of hypothetical similar studies.

Another approach is to define probability not as a trait of data or nature, but as a property of a spectator. This means that probability is seen as a degree of belief, potentially different for everyone. Experience is then seen as a way of updating these ideas of reality: each time we experience an event we revise our belief about what caused this event in light of that particular experience. Formalization of this process in science requires usage of some probability mathematics called 'Bayes' rule', and the statistical methods are therefore often referred to as being part of the 'Bayesian framework'. Because Bayesian analyses estimate underlying parameters directly, there is no need for *P* values and the interpretation of confidence intervals is much more straightforward. The latter refer directly to parameters.

Although the Bayesian framework is highly compatible with the human way of thinking and acting, science has for long revolted against introducing obvious and explicit subjectivity in its realm[14]. Apart from the philosophical debate, Bayesian analyses have been relatively uncommon due to the lack of computing power that is required for solving the high-dimensional integrals in the calculations. With modern computers and simulation-based approaches this is no longer a major issue and nowadays both Frequentist and Bayesian approaches are being used together for a variety of analyses. In this dissertation, primarily Frequentist methods are used, where appropriate complemented by Bayesian approaches.

1.4 Why meta-analysis and why not

There are a number of reasons why it could be preferable to do a meta-analysis instead of a qualitative description of previous research results. First of all, virtually all primary studies report their results numerically and 'hard' numerical results is what most patients, physicians, and other decision makers in health care would like to base their decisions on. It is a waste to lose this quantitative dimension once there is more than one study performed on a certain topic.

Another point that is often made in favor of meta-analysis is that it combines the samples of multiple trials and thus increases the precision of the estimations. This is also illustrated in Figure 1.1, where the width of diamond in the bottom right plot indicates the confidence interval of the meta-analysis summary estimate (which is substantially smaller than the confidence intervals of the individual studies).

Although extensive differences between studies are a reason not to do a meta-analysis[8-10], if the differences are not too great and the studies are still estimating a similar parameter, it can be argued that combining studies increases generalizability of the estimate over a larger population[15]. Even if the studies are indeed too different to generalize meaningfully to a population, meta-analytical techniques such as subgroup analyses and (hierarchical) meta-regression can be used to explore the reason for variation between studies[16-18].

Finally, meta-analysis can be used to explore trends in the accumulation of evidence (cumulative meta-analysis[19]) and assess the impact of information from a subsequent study. This way, meta-analysis can also guide decisions on whether additional research needs to be performed and under what circumstances (e.g. with what observed result) it would contribute to or change existing opinions on a topic. Although there are many reasons why a meta-analysis would be indicated, there are also reasons why one might refrain from doing one. First of all, if studies are

too different in too many respects, combining them produces results that make no sense in any of the domains of each included study. It may also make no sense to do a meta-analysis in a field where developments go so fast that interventions and technologies are replaced by new ones before a meaningful systematic review of existing evaluative research reports can be done.

Another type of problem arises when primary studies included in the meta-analysis are biased or when they are only a subset of all the studies that were done on the topic. The latter is often referred to as dissemination or publication bias. Although more precise estimates are generally to be preferred over less precise estimates, it is well appreciated that a precise biased estimate is worse than an imprecise biased estimate or no numerical estimate at all.

The above-mentioned issues underline the need to perform meta-analyses only in the context of a systematic review in which potential biases are systematically prevented or assessed. When we talk about meta-analysis in this dissertation, we therefore assume that it is performed as part of a systematic review.

1.5 Aim and outline of this dissertation

The general aim of the projects described in this dissertation was to develop and evaluate meta-analysis methods with applications to medical science. The dissertation is divided in five major parts: Introduction, Methodology, Applications, Discussion and Summary, and Appendices. The Methodology part starts with Chapter 2, describing an investigation into the level of experience of contemporary authors of meta-analyses. Chapter 3 analyzes the terminology related to dissemination bias, and Chapter 4 reports a comprehensive evaluation of the use of graphs in meta-analysis. Chapter 5 and Chapter 6 describe the development of statistical software for meta-analysis, and report a comparative review of meta-analysis software, respectively. In the Applications part, Chapter 7 and Chapter 8 report two meta-analyses in which methods and software described in the previous chapters are applied. The Discussion and Summary section starts with a comprehensive paper on methods of meta-analysis. This is followed by a summary in English and in Dutch. The dissertation ends with a number of appendices, acknowledgments, and a short biography of the author.

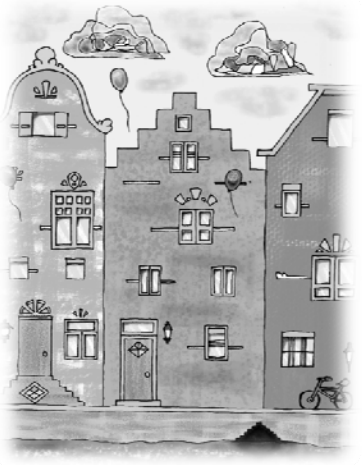
References

1. Glass GV. Primary, secondary and meta-analysis research. *Educ Res.* 1976;5:3-8.

2. Pearson K. Report on certain enteric fever inoculation statistics. *BMJ.* 1904;3:1243-1246.

3. Birge RT. The calculation of errors by the method of least squares. *Phys Rev.* 1932;16(1-32).
4. Cochran WG. Problems arising in the analysis of a series of similar experiments. *J Roy Stat Soc.* 1937;4(Supplement):102-118.
5. Light RJ, Smith PV. Accumulating evidence: procedures for resolving contradictions among different research studies. *Harvard Educ Rev.* 1971;41:429-471.
6. Hedges LV, Olkin I. Vote-counting methods in research synthesis. *Psychol Bull.* 1980;88:359-369.
7. Greenland S. Quantitative methods in the review of epidemiologic literature. *Epidemiol Rev.* 1987;9:1-30.
8. Egger M, Davey Smith G, Altman D. *Systematic Reviews in Health Care: Meta-Analysis in Context.* London: BMJ Publishing Group; 2001.
9. Glasziou P, Irwig L, Bain C, Colditz G. *Systematic Reviews in Health Care: A Practical Guide*. Cambridge: Cambridge University Press; 2001.
10. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. *Methods for Meta-Analysis in Medical Research*. Chichester: Wiley; 2000.
11. Miettinen OS, Flegel KM. Elementary concepts of medicine: VIII. Knowing about a client's health: gnosis. *J Eval Clin Pract.* Aug 2003;9(3):333-335.
12. Miettinen OS. Knowledge base of scientific gnosis: II. Gnostic occurrence relations: elements and temporal structures. *J Eval Clin Pract.* May 2004;10(2):357-359.
13. Miettinen OS. Knowledge base of scientific gnosis: III. Gnostic occurrence relations as regression functions. *J Eval Clin Pract.* May 2004;10(2):361-363.
14. Press JS, Tanur JM. *The Subjectivity of Scientists and the Bayesian Approach*. New York: Wiley; 2001.
15. Collins R, Gray R, Godwin J, Peto R. Avoidance of large biases and large random errors in the assessment of moderate treatment effects: the need for systematic overviews. *Stat Med.* Apr-May 1987;6(3):245-254.
16. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. *Stat Med.* Apr 30 1998;17(8):841-856.
17. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ.* Nov 19 1994;309(6965):1351-1355.
18. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Stat Med.* Oct 30 1999;18(20):2693-2708.
19. 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.* Jul 23 1992;327(4):248-254.

METHODOLOGY



2

WHO EXPLORES AND SUMMARIZES OUR EVIDENCE?

Based on:

Bax L, Tsuneda N, Wang G, Satoh T, Moons KG.

Who explores and summarizes our evidence?

Submitted for publication.

Abstract

Introduction. Meta-analyses are a pivotal part of systematic reviews and a cornerstone of scientific medicine. We often assume that the people responsible for the meta-analyses are experienced researchers, but is this really true?

Methods. We retrieved all citations in Medline (PubMed) that were classified as a meta-analysis in 2008. We subsequently performed a systematic exploration of other Medline-registered meta-analyses of the first and second or last authors.

Results. Roughly 41% of the first authors had been involved in a meta-analysis, and 28% as primary investigator. The second and last authors combined were commonly more experienced, although 32% had never participated in a meta-analysis and 59% had not been involved in one as a first author.

Conclusion. Primary investigators of meta-analyses are often not experts in meta-analysis. Co-authors were generally more experienced. This has implications for how meta-analyses are (re)viewed and how educational materials and software for meta-analysts are designed.

2.1 Introduction

Meta-analyses are a pivotal part of systematic reviews and a cornerstone of scientific medicine[1]. In medicine, the motive to conduct a systematic review and meta-analysis is often to inventory the existing clinical evidence on a certain topic. The summary results are often applied in clinical guidelines or used to guide the development of new research projects.

Before the popularization of systematic reviews in the 1990s, narrative reviews were primarily invited papers, written by experts in the field[2]. When in the 1980s and 1990s – supported by the introduction of evidence-based medicine – medical decision-making became more quantitative, many came to realize that narrative reviews were not providing optimal input for medical practice. These developments led to the popularization of systematic reviews and the inauguration of the Cochrane Collaboration in October 1993. In order to be transparent and reproducible, systematic reviews follow explicit protocols similar to those applied in clinical trials[4]. Although this places review studies in the realm of science, the explicitly described methods and consequent reproducibility means reviewing is no longer the sole domain of well-known experts. Moreover, if subjective interpretations are banned from a review, being an expert with an opinion may even become a handicap to the review author.

The number of publications classified as meta-analysis in PubMed has grown exponentially over the years (Figure 2.1). We hypothesized that the majority of systematic reviews are performed by relatively inexperienced researchers and set out to investigate the level of experience of the people that are responsible for modern-day meta-analyses.

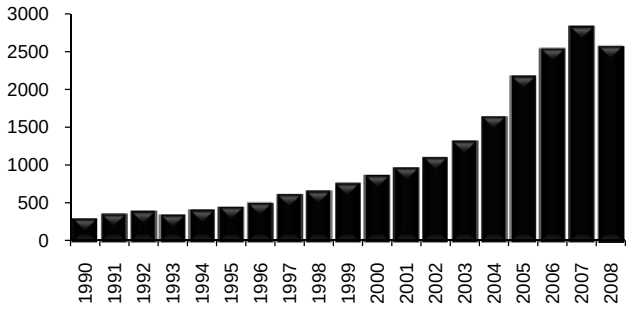


Figure 2.1 Meta-analyses registered in PubMed since 1990

An overview of published meta-analyses registered in PubMed between 1990 and 2008. The graph is based on data from a search in February 2009, using "meta-analysis"[pt] AND YYYY[dp] as search filter.

2.2 Methods

2.2.1 Design

Our objective was to make a simple inventory of how experienced or inexperienced current principal investigators (first authors) of systematic reviews and meta-analyses are. For reasons of convenience and temporal acuity we assessed all meta-analyses of the year 2008. We used Medline (PubMed) as resource database, as it indexes meta-analyses with a publication-type tag. Experience was quantified by the number of meta-analyses and number of publications the first authors had (co-)authored before their 2008 meta-analysis publication(s).

2.2.2 Search and selection

We retrieved all citations in PubMed that were classified as meta-analysis of human studies in 2008 (meta-analysis[pt] AND 2008[dp] AND humans[mh]). The retrieved citations were subsequently imported in Excel. Citations with only electronic publication in 2008 and actual publication in 2009 as well as double citations of the same article were deleted. Finally, the data set was reshaped with authors instead of citations as units of analysis. For authors with more than one meta-analysis citation in 2008, the most recent one was kept in the data set and the others were counted as previous experience.

2.2.3 Data extraction and author citation search

For each first author of the retrieved 2008 meta-analyses, we first documented the number of PubMed-registered papers that he or she had published (as first author or co-author) before the 2008 meta-analysis. However, because PubMed does not give unique identifiers to authors, we anticipated that some authors with common names would be classified as having published an unlikely high number of articles. Because including the data from such authors biases the measured experience in the full data set upward, we decided to omit the authors with more than 300 papers. We performed sensitivity analyses for lower cut-off points like 200 and 100 articles.

Although we checked the number of published papers in general for selection purposes, our primary focus was on how many of the authors had previously been involved in meta-analyses. We also investigated the number of meta-analyses previously published by the article's second or last authors.

2.2.4 Data analysis

We calculated descriptive statistics for the number of previous papers in Microsoft Excel. Because the research question was not comparative, no significance tests or formal comparisons were performed.

2.3 Results

On January 16, 2009, we found a total of 2320 citations registered as meta-analysis in PubMed in 2008. After deleting double citations, citations without an author tag, and citations without actual publication date in 2008, 1983 citations remained available for further analysis. After restructuring the data set, the final analysis data frame contained 1838 meta-analyses from unique authors.

Approximately 28% of the first authors of the meta-analyses in 2008 had been the first author of a meta-analysis before and 41% had been involved in a meta-analysis as first or co-author. Roughly 14% had published two or more meta-analyses as first author and 24% as first or co-author. In negative terms, 72% of the meta-analyses' first authors had not been involved in any meta-analysis as first author and 59% had not had any meta-analysis experience before 2008. The results are summarized in the left column of Table 2.1.

Table 2.1 Prior meta-analysis experience of the included meta-analyses' authors

	2008 meta-analyses' first authors	2008 meta-analyses' second / last authors
Previous meta-analysis experience as first author		
0 papers (%)	72.1%	58.7%
1 paper (%)	14.7%	16.1%
2 papers (%)	5.6%	8.8%
>=3 papers (%)	7.6%	16.3%
Previous meta-analysis experience as co-author		
0 papers (%)	58.9%	32.0%
1 paper (%)	17.0%	14.5%
2 papers (%)	7.3%	9.6%
>=3 papers (%)	16.8%	43.8%

The table describes the experience of the authors of the 2008 meta-analyses (k=1838). gives the percentage of authors that published 0, 1, 2, and 3 or more previous meta-analyses. The data are given for the first authors of the 2008 meta-analyses (left column) and for the second or last authors of the 2008 meta-analyses (right column).

The second and last authors together (right column of Table 2.1) commonly brought in more experience than the first authors. Still, approximately 59% had not written a meta-analysis paper as first author. About 32% of the co-authors had

not collaborated in a meta-analysis (as first or co-author) before their 2008 meta-analysis, but 44% had (co-)authored three or more. We also found that meta-analysis is indeed teamwork, with 81% of the 2008 meta-analyses having three or more authors and 62% having four or more.

Sensitivity analyses indicated that the results were robust to changing the cut-off for exclusion of authors with an unlikely high number of published papers from 300 to 200 or 100 citations. Lowering the cut-off excluded more authors and slightly decreased the previous experience attributed to the remaining authors (data available upon request).

2.4 Discussion

We systematically reviewed the citations registered as meta-analysis in PubMed in 2008 and found that as little as 41% of the first authors of these meta-analyses had been involved in a meta-analysis before and only 28% as primary investigator. The second and last authors combined were more experienced, with approximately 68% having worked on a meta-analysis and 41% as primary investigator.

We are not aware of investigations similar to the one we did, and it is therefore hard to provide much more than speculation about how this information fits in current beliefs about meta-analysis and meta-analysts. Although many may assume that meta-analysts are well-established researchers, we believe there is a trend among young researchers to perform formal meta-analyses before they embark on a new primary research project. This is a positive trend and it naturally fits in the idea that meta-analysis can be used to quantify the potential value of new research projects and subsequent evidence.

A few methodological limitations should be discussed. First, the indexing process of meta-analyses in PubMed is not 100% accurate. The misclassification works in both directions: some meta-analyses are not classified as such while some non-meta-analytical studies are tagged with the meta-analysis publication type. It is unclear which type of misclassification is most common and whether the studies we attributed to the meta-analysis experience of authors are an over-retrieval or under-retrieval.

Another issue in PubMed is that there are no unique identifiers for authors. Hence, we inevitably ended up with too many papers and too many meta-analyses attributed to what appeared to be a single author. We discarded authors with common names and an unlikely number of published papers (>300), but we believe the results from our assessment may still be biased upwards (attributing

too much experience to the authors). Sensitivity analyses with lower cut-off points resulted in somewhat lower estimates of prior experience.

We did not check how many of the prior meta-analysis citations were reports of the same project and this could also have caused an upward bias. We found that it was quite common to find multiple meta-analysis publications on the same topic and in the same year from the same first author. In some cases, the extra citations were due to a flawed indexing process (citations identical except for a volume number). However, in many cases the titles were identical except for an outcome (morbidity instead of mortality) or a different treatment (different drug but with similar therapeutic mechanisms). For example, Bjelakovic et al. published a Cochrane Review titled "Antioxidant supplements for prevention of mortality in healthy participants and patients with various diseases"[5] and at the same time also published a Cochrane Review titled "Antioxidant supplements for preventing gastrointestinal cancers"[6]. The latter is clearly aimed at identical populations but reporting a different, near-surrogate outcome of the former.

Our finding that researchers with relatively little experience in meta-analysis contribute to a large part of meta-analytical evidence has a number of implications. First, it indicates that systematic reviews and meta-analyses must be teamwork, preferably involving (experienced) clinicians as well as statisticians or clinical epidemiologists. Second, it implies that good peer-review is essential to ensure high-quality meta-analysis publications. Complementary narrative reviews may also be important to provide an empirical reality check for the quantitative and at times narrow approach in systematic reviews[3]. Furthermore, the complex statistical advances in meta-analysis that have been made over the last ten years are to some extent in contrast with the level of experience of those that are to apply these methods. Easy-to-use software and non-technical articles may be instrumental in bringing such methods to a larger audience.

As a final note, we do not want to suggest that a lack of experience of authors necessarily invalidates the results of a properly executed meta-analysis. However, since reviews and meta-analyses are often pivotal in clinical guideline development and decision-making, we do believe that it is important to subject reviews to a thorough re-view, preferably by experts in clinical medicine as well as experts in meta-analysis.

Conflict of interest

None declared.

References

1. Eggers M, Davey Smith G, Altman DG. *Systematic Reviews in Health Care: Meta-Analysis in Context.* London: BMJ Publishing Group; 2001.

2. Chalmers I, Hedges LV, Cooper H. A brief history of research synthesis. *Eval Health Prof.* Mar 2002;25(1):12-37.

3. Collins JA, Fauser BC. Balancing the strengths of systematic and narrative reviews. *Hum Reprod Update.* Mar-Apr 2005;11(2):103-104.

4. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.1: The Cochrane Collaboration. *www.cochrane-handbook.org* (updated September 2008).

5. Bjelakovic G, Nikolova D, Gluud LL, Simonetti RG, Gluud C. Antioxidant supplements for prevention of mortality in healthy participants and patients with various diseases. *Cochrane Database Syst Rev.* 2008(2):CD007176.

6. Bjelakovic G, Nikolova D, Simonetti RG, Gluud C. Antioxidant supplements for preventing gastrointestinal cancers. *Cochrane Database Syst Rev.* 2008(3):CD004183.

3

PUBLICATION BIAS: A PARS PRO TOTO

Based on:

Bax L, Moons KG.

Publication bias: A pars pro toto.

Submitted for publication.

Abstract

Introduction. The term publication bias is widely used to describe the tendency of investigators, reviewers, and editors to submit or accept manuscripts for publication based on the direction or strength of the study findings. It is not commonly recognized that the term is inaccurate.

Methods. In an attempt to provide clarity and consistency in terminology, we provide an essay about selectivities and biases in the dissemination of evidence.

Results. Evidence dissemination can be viewed as a multi-staged process with dissemination alternatives at distinct points in time. Selectivities in reporting (of outcomes as well as entire studies), publication, and inclusion of evidence can lead to a number of biases. We propose to use the terms reporting bias, publication bias, and inclusion bias. The aggregate of these biases is in our view best described by the term dissemination bias. The common usage of the term publication bias to represent all biases thus constitutes a 'pars pro toto' – a part to describe the whole – and is inappropriate. We also distinguish meta-analytical methods that make a judgment on whether dissemination selectivity has likely occurred or not, and methods that attempt to quantify or correct for the bias that may have been induced by the dissemination selectivity.

Conclusion. Terminology in the dissemination of evidence has been ambiguous. A careful use of terms can facilitate the understanding of the problems that occur in the dissemination of evidence.

3.1 Introduction

Searching the Internet for "publication bias" with Google leads to an overwhelming number of hits. Searching Medline via PubMed with ("publication bias"[mh] OR "publication bias"[ti] OR "small-study effect*"[ti] OR "dissemination bias"[ti]) yields roughly 1180 citations that refer to the topic (February 2009). An early paper in the social and statistical sciences about effects of publication decisions appeared in the Journal of the American Statistical Association in 1959[1]. The first reference that explicitly names and defines publication bias in the medical literature is a paper by Begg[2] in 1985, which states that "studies in which the observed efficacy of the treatment is high are much more likely to be reported than those in which the observed efficacy is average or poor… the reported efficacy of a treatment might be inflated … this quantity is termed the publication bias". The following year, Simes[3] used the term in a paper entitled "Publication bias: the case for an international registry of clinical trials", and a number of methodological papers dedicated to this topic appeared soon thereafter[4-9]. The interest for the topic has not waned over the years, and recent publications include statistical papers[10-15] as well as reviews and theoretical work[16-19].

That the term publication bias is misleading or at least only describing part of the problem was suggested by Chalmers in 1990[4], when he described post-publication biases and pre-publication biases[4]. It was not until ten years later that Fujian Song deliberated on this in a paper called "Publication and related biases"[20]. The paper suggests the use of the term 'dissemination bias' instead of publication bias. The Cochrane Collaboration's online learning materials[21], on the other hand, suggest reporting bias as summary term to describe this group of biases. Other authors reserve this term for the bias due to under or over-reporting of study results by authors and do not use this term to describe bias due to selective rejection of submitted papers by journal editors[22-24].

A recent book titled "Publication bias in meta-analysis"[25] explicitly acknowledges that the term is unfortunate and inaccurate, but states that "the established use of the term publication bias has made us hesitant to tamper with, and potentially confuse, the current terminology". This statement, however, ignores the fact there is already confusion, as shown above. Although a continued use of the established terminology may serve the incidental reader, this is perhaps not the correct approach when teaching a new generation of medical doctors and researchers. We therefore provide a short essay on (alternative) terminology in the dissemination of (medical research) evidence, with special reference to meta-analysis.

3.2 Distinguishing selectivity and bias

We define evidence dissemination as the process by which scientific data are transformed into evidence for the public domain. This starts when study results are summarized by researchers and ends when they reach the public directly by presentations or published reports of the primary study or as part of a systematic review.

The word bias originally meant oblique and is still used in that meaning by tailors (a line going diagonally across the grain of fabric). In common language it has come to refer to a 'tendency' or 'prejudice'[26]. In statistics, however, bias is defined as the deviation of a statistical estimate from the quantity it estimates or systematic error in the study results[27]. We adopt the latter way of thinking and further distinguish systematic error in the study results from systematic error in the study design or measurement process. This distinction is necessary, we believe, because the two errors reflect different concepts and are not necessarily related: a methodologically flawed process does not lead to biased results per se. For example, the randomization for a clinical trial may not have been performed properly, resulting in a selective (unequal) distribution of some prognostic factors over the study arms. Still, the results may not be biased. In derived or meta-analytical research, a systematic review search filter may have missed a number of studies, but the number of missing studies may be small or distributed such that there is no influence or bias in the meta-analysis's results. In the context of evidence dissemination, we will refer to the process error as selectivity and will refer to the results error as bias.

3.3 The stages of evidence dissemination

Evidence dissemination can be viewed as a multi-staged process with dissemination alternatives at distinct points in time (Figure 3.1). At each point in time a selective process can occur that determines which results get disseminated and which results are suppressed. We propose the term 'dissemination selectivity' to refer to selective distribution or censoring in the dissemination process. The selectivity can be positive or negative; negative selectivity can be regarded as the censoring of study results and positive selectivity as the over-representation of study results. 'Reporting selectivity' occurs if researchers themselves decide to submit only part of their data (e.g. with a significant result) to a journal (outcome reporting selectivity) or perhaps not submit their results at all (study reporting selectivity). At a later stage, publication selectivity occurs if editors or referees of a journal decide not to accept certain manuscripts for publication, notably those with non-significant, negative results. Finally, if a study is reported and published

but not included in a database or (systematic) literature review, this can be referred to as 'inclusion selectivity', specifically database inclusion selectivity (e.g. selective inclusion of certain studies in Medline) and review inclusion selectivity (e.g. selective inclusion of studies in reviews).

Naturally, each form of dissemination selectivity can lead to bias: outcome reporting bias, study reporting bias, publication bias, and inclusion bias. All of these biases can be viewed as being part of a larger construct called dissemination bias. Numerous other types of biases have been described in relation to the dissemination of evidence (e.g. language bias[28,29], funding bias[30,31], database bias[32], citation bias[33,34], location bias[35,36], reviewer bias[37]). In our terminology, one can specify at what stage the selective dissemination occurs that causes these biases. For example, language bias may occur due to a combination of selective reporting (researchers only taking the time to write in English if the results of their study are significant and positive) as well as selective inclusion (databases and systematic reviews only including English-language journals and papers). Funding bias, on the other hand, may primarily be due to selective reporting (researchers not allowed by their funding source to report certain results to journals).

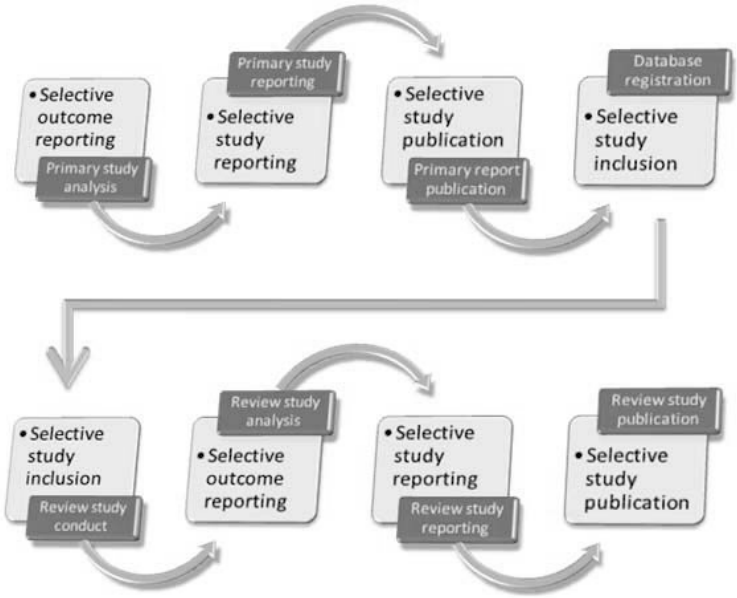


Figure 3.1 The sequential process of evidence dissemination

Evidence is disseminated in sequential steps during which study authors, journal editors, database editors, and review authors (in chronological order) make selections as to which evidence is disseminated and which evidence is discarded.

The dissemination selectivities are empirical issues that require (separate) preventive strategies and it is therefore possible and useful to address them individually (upper part of Figure 3.2). However, when exploring and summarizing evidence in a meta-analysis, only overall trends in the data at hand (e.g. an under-representation of study results) can be distinguished. One cannot reliably attribute the trends to individual (dissemination) selectivities (lower part of Figure 3.2). The trends are reflections of selectivity in general, possibly occurring by chance or unrelated to the dissemination process[38-41]. Although it is common to speak of publication bias assessments in a meta-analysis[16,42-46], we believe this terminology is incorrect and confusing. Because the selectivities cannot be distinguished, neither can their resulting biases.

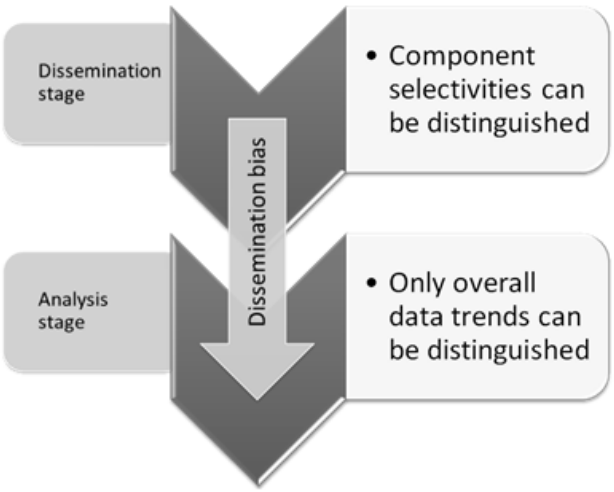


Figure 3.2 Bias from the dissemination stage to the analysis stage

The various component selectivities that lead to dissemination bias in primary studies cannot be disentangled in a meta-analysis. Instead, assessments focus on data trends that are caused by unknown selectivity processes.

3.4 Analysis of data trends and their impact

A variety of tests have been designed to determine whether the evidence in the meta-analysis is likely to be representative of the full ecology of the available research data. In line with the distinction between selectivity and bias, we believe that in meta-analysis a distinction should be made between methods that attempt to quantify data trends due to selectivity and those that attempt to quantify the bias resulting from this selectivity.

The common rationale behind meta-analytical assessments of data trends in relation to dissemination selectivity is that if selective under-reporting, under-

publishing, or under-inclusion has occurred it will most likely affect the small and non-significant negative studies that were not seen as interesting. From this assumption it follows that if the effect size (e.g. the risk ratio, odds ratio, or a mean difference) is plotted against the precision (inverse of the standard error of the effect), dissemination selectivity could cause small studies with low precision and small (negative) effects to appear to be missing (the half-circle area in Figure 3.3a).

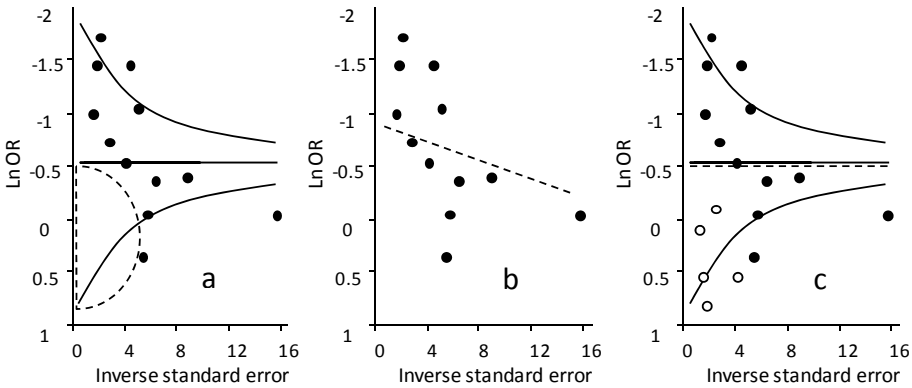


Figure 3.3 Funnel plots

Three funnel plots, made for the purpose of illustration. The effect size (here the logarithm of the odds ratio, for example measuring the effect of a particular treatment for a particular disease) is plotted against the precision of the effect size (inverse standard error).The summary estimate of the individual odds ratios is depicted by a horizontal solid line. The left funnel plot (a) also contains pseudo-confidence limits (funnel-shaped lines) that can be helpful in judging symmetry around the summary effect and a half-circle area in which small negative studies are expected to be found. The funnel plot in the center (b) shows a regression plot with a weighted regression line (dashed down-sloped line). Missing small studies can induce an association between the effect size and the precision, as is shown by the regression line. The right plot (c) shows the results of the trim-and-fill algorithm with studies imputed (empty circles) to correct for the void in the half-circle area in the plot on the far left. As a result the meta-analysis summary estimate of the log odds ratio has shifted towards zero (shown by the dashed horizontal line).

The plot is called a funnel plot and the data trend is also referred to as small-study effects[45,47]. A number of regression approaches[47-50] exist to quantify the described data trends. Depending on the approach, a significant *P* value of the intercept or slope of the (weighted) regression line is thought to indicate data trends that could be due to dissemination selectivity (Figure 3.3b). Rank correlation methods[51,52] examine the association between the effect size and the precision in a non-parametric way.

It is also possible to go beyond the dichotomous judgment of presence of selectivity and assess the bias resulting from the selectivity or even correct for it.

A relatively simple method that is used for this purpose is the trim-and-fill approach[53]. The trim-and-fill method assumes selectivity may cause an imbalance in the ranks of the effect sizes of positive and negative studies. If an imbalance is detected, a number of studies are imputed to restore the balance. The imputed studies often fill the gap in the funnel plot where small studies appear to be missing (Figure 3.3c), or they compensate for the missingness in that area. Some regression approaches[15,47-50,52,54,55] can also be extended to provide a meta-analysis result that is corrected for the possible dissemination selectivity underlying the data.

The tests for data trends that may be due to dissemination selectivity and the quantifications of subsequent bias should be seen as complementary. The former provides information about the likelihood that selective processes may have had an influence on the data occurrence and the latter about the consequences these processes may have on the validity of the results.

3.5 Conclusion

To base decisions on evidence that has been subject to selectivity in the process of dissemination of evidence poses a major threat to the validity of one's conclusions. We distinguish a number of selective processes in reporting (of outcomes as well as entire studies), publication, and inclusion (in databases and reviews) of evidence. These selectivities lead to bias, i.e. reporting bias, publication bias, and inclusion bias. In line with Song et al.[20], we propose the term dissemination bias to represent this family of biases. The common usage of the term publication bias to represent all biases constitutes a 'pars pro toto' and is inappropriate. In line with the distinction between selectivities and their resulting biases, we distinguish meta-analytical methods that make a judgment on whether dissemination selectivity has likely occurred or not and meta-analytical methods that attempt to quantify or correct for the bias potentially induced by the dissemination selectivity. Both types of methods have a place in meta-analysis and should be seen as complementary.

Conflict of interest

None declared.

References

1. Sterling TD. Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa. *J Am Stat Assoc.* 1959;54:30-34.

2. Begg CB. A measure to aid in the interpretation of published clinical trials. *Stat Med.* Jan-Mar 1985;4(1):1-9.

3. Simes RJ. Publication bias: the case for an international registry of clinical trials. *J Clin Oncol.* Oct 1986;4(10):1529-1541.

4. Chalmers TC, Frank CS, Reitman D. Minimizing the three stages of publication bias. *JAMA.* Mar 9 1990;263(10):1392-1395.

5. Chalmers TC, Levin H, Sacks HS, Reitman D, Berrier J, Nagalingam R. Meta-analysis of clinical trials as a scientific discipline. I: Control of bias and comparison with large co-operative trials. *Stat Med.* Apr-May 1987;6(3):315-328.

6. Dickersin K. The existence of publication bias and risk factors for its occurrence. *JAMA.* Mar 9 1990;263(10):1385-1389.

7. Dickersin K, Chan S, Chalmers TC, Sacks HS, Smith H, Jr. Publication bias and clinical trials. *Control Clin Trials.* Dec 1987;8(4):343-353.

8. Newcombe RG. Towards a reduction in publication bias. *Br Med J (Clin Res Ed).* Sep 12 1987;295(6599):656-659.

9. Vandenbroucke JP. Passive smoking and lung cancer: a publication bias? *Br Med J (Clin Res Ed).* Feb 6 1988;296(6619):391-392.

10. Moreno SG, Sutton AJ, Ades AE, et al. Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Med Res Methodol.* Jan 12 2009;9(1):2.

11. 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.

12. Saveleva E, Selinski S. Meta-analyses with binary outcomes: how many studies need to be omitted to detect a publication bias? *J Toxicol Environ Health A.* 2008;71(13-14):845-850.

13. Rucker G, Schwarzer G, Carpenter J. Arcsine test for publication bias in meta-analyses with binary outcomes. *Stat Med.* Feb 28 2008;27(5):746-763.

14. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *J Clin Epidemiol.* Oct 2008;61(10):991-996.

15. Copas JB, Malley PF. A robust P-value for treatment effect in meta-analysis with publication bias. *Stat Med.* Sep 20 2008;27(21):4267-4278.

16. Ioannidis JP. Why most published research findings are false. *PLoS Med.* Aug 2005;2(8):e124.

17. Young NS, Ioannidis JP, Al-Ubaydli O. Why current publication practices may distort science. *PLoS Med.* Oct 7 2008;5(10):e201.

18. Lenzer J, Brownlee S. An untold story? *BMJ.* Mar 8 2008;336(7643):532-534.

19. Kheifets L, Olsen J. Should epidemiologists always publish their results? Yes, almost always. *Epidemiology.* Jul 2008;19(4):532-533.

20. Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ. Publication and related biases. *Health Technol Assess.* 2000;4(10):1-115.

21. Green S, Anderson P. Cochrane Collaboration open learning material for reviewers. The Cochrane Collaboration. *http://www.cochrane-net.org/openlearning/html/mod15-2.htm*. Accessed April 2009.

22. Dwan K, Altman DG, Arnaiz JA, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS ONE.* 2008;3(8):e3081.

23. Rifai N, Altman DG, Bossuyt PM. Reporting bias in diagnostic and prognostic studies: time for action. *Clin Chem.* Jul 2008;54(7):1101-1103.

24. 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.

25. Rothstein HR, Sutton AJ, Borenstein M. *Publication bias in meta-analysis. Prevention, assessment, and adjustments.* Chichester: John Wiley & Sons; 2005.

26. Merriam-Webster Online. Merriam-Webster Online Dictionary. *http://www.merriam-webster.com*.

27. Upton G, Cook I. *A Dictionary of Statistics*. New York: Oxford University Press; 2008.

28. Egger M, Zellweger-Zahner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomised controlled trials published in English and German. *Lancet.* Aug 2 1997;350(9074):326-329.

29. Juni P, Holenstein F, Sterne J, Bartlett C, Egger M. Direction and impact of language bias in meta-analyses of controlled trials: empirical study. *Int J Epidemiol.* Feb 2002;31(1):115-123.

30. Davidson RA. Source of funding and outcome of clinical trials. *J Gen Intern Med.* May-Jun 1986;1(3):155-158.

31. Bero L, Oostvogel F, Bacchetti P, Lee K. Factors associated with findings of published trials of drug-drug comparisons: why some statins appear more efficacious than others. *PLoS Med.* Jun 2007;4(6):e184.

32. Zielinski C. New equities of information in an electronic age. *BMJ.* Jun 10 1995;310(6993):1480-1481.

33. Kjaergard LL, Gluud C. Citation bias of hepato-biliary randomized clinical trials. *J Clin Epidemiol.* Apr 2002;55(4):407-410.

34. Chapman S, Ragg M, McGeechan K. Citation bias in reported smoking prevalence in people with schizophrenia. *Aus NZ J Psych.* 2009;43(3):277 - 282.

35. Vickers A, Goyal N, Harland R, Rees R. Do certain countries produce only positive results? A systematic review of controlled trials. *Control Clin Trials.* Apr 1998;19(2):159-166.

36. Pittler MH, Abbot NC, Harkness EF, Ernst E. Location bias in controlled clinical trials of complementary/alternative therapies. *J Clin Epidemiol.* May 2000;53(5):485-489.

37. Ernst E, Resch KL. Reviewer bias: a blinded experimental study. *J Lab Clin Med.* Aug 1994;124(2):178-182.

38. Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet.* Aug 22 1998;352(9128):609-613.

39. Schulz KF. Subverting randomization in controlled trials. *JAMA.* Nov 8 1995;274(18):1456-1458.

40. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA.* Feb 1 1995;273(5):408-412.

41. Sterne JA, Juni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. *Stat Med.* Jun 15 2002;21(11):1513-1524.

42. Rothstein HR, Sutton AJ, Borenstein M. *Publication Bias in Meta-Analysis. Prevention, Assessment, and Adjustments.* Chichester: John Wiley & Sons; 2005.

43. Egger M, Davey Smith G, Altman DG. *Systematic Reviews in Health Care: Meta-Analysis in Context.* London: BMJ Publishing Group; 2001.

44. Mahid SS, Qadan M, Hornung CA, Galandiuk S. Assessment of publication bias for the surgeon scientist. *Br J Surg.* Aug 2008;95(8):943-949.

45. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *J Clin Epidemiol.* Nov 2000;53(11):1119-1129.

46. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. *Methods for Meta-Analysis in Medical Research.* Chichester: Wiley; 2000.

47. Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Stat Med.* Oct 30 2006;25(20):3443-3457.

48. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ.* Sep 13 1997;315(7109):629-634.

49. Macaskill P, Walter SD, Irwig L. A comparison of methods to detect publication bias in meta-analysis. *Stat Med.* Feb 28 2001;20(4):641-654.

50. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Comparison of two methods to detect publication bias in meta-analysis. *JAMA.* Feb 8 2006;295(6):676-680.

51. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics.* Dec 1994;50(4):1088-1101.

52. Schwarzer G, Antes G, Schumacher M. A test for publication bias in meta-analysis with sparse binary data. *Stat Med.* Feb 20 2007;26(4):721-733.

53. Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics.* Jun 2000;56(2):455-463.

54. Copas J, Jackson D. A bound for publication bias based on the fraction of unpublished studies. *Biometrics.* Mar 2004;60(1):146-153.

55. Copas JB, Shi JQ. A sensitivity analysis for publication bias in systematic reviews. *Stat Methods Med Res.* Aug 2001;10(4):251-265.

4

More Than Numbers: The Power of Graphs in Meta-Analysis

Based on:
Bax L, Ikeda N, Fukui N, Yaju Y, Tsuruta H, Moons KG.
More Than Numbers: The Power of Graphs in Meta-analysis
Am J Epidemiol. 2009; 169: 249-255.

Abstract

Introduction. In meta-analysis, the assessment of graphs is widely used in an attempt to identify or rule out heterogeneity and publication bias. A variety of graphs are available for this purpose. To date, however, there has been no comparative evaluation of the performance of these graphs.

Methods. Our objective was to assess the reproducibility and validity of the graph ratings. We simulated 100 meta-analyses from four scenarios that covered situations with and without heterogeneity and publication bias. From each meta-analysis we produced 11 graphs (box plot, weighted box plot, standardized residual histogram, normal quantile plot, forest plot, three funnel plots, trim-and-fill plot, Galbraith plot, and L'Abbé plot) and assessed the resulting 1100 plots with three reviewers.

Results. The intra-class correlation coefficients for reproducibility of the graph ratings ranged from poor (0.34) to high (0.91). Ratings of the forest plot and the standardized residual histogram were best associated with parameter heterogeneity. Association between graph ratings and publication bias (censorship of studies) was poor, with the weighted box plot and the trim-and-fill plot showing the best combination of reproducibility and validity.

Conclusion. Meta-analysts should be selective in the graphs they choose for the exploration of their data.

4.1 Introduction

Over the past decades, systematic reviews - with meta-analyses as their quantitative and analytical core - have become a cornerstone of evidence-based medicine. As such, meta-analyses play a central role in the development of clinical guidelines and in clinical decision-making. Most meta-analytical endeavors involve graphs, mostly forest or funnel plots, to decide which analytical approach serves the synthesis of the study data best, but notably to explore heterogeneity and possible publication bias[1-4]. Heterogeneity and publication bias are perhaps the two major threats to the validity of a meta-analysis.

Heterogeneity refers to the variation among effect parameters across studies. In meta-analyses one commonly tests whether the assumption of a single underlying effect (fixed effect analysis) or similar effects (random effects analysis) is indeed reasonable. Approaches based on the assumption of a fixed effect can be used if the between-study variation is not excessive. Approaches based on the assumption of random effects are used to incorporate the heterogeneity if present. In the latter case, sources of heterogeneity can be investigated and incorporated by stratified meta-analyses or meta-regression[5].

Publication bias can be viewed as a systematic error in a meta-analysis that occurs because not all evidence is properly represented. Although there is both over and under-representation of evidence, under-representation is more prominent and often refers to 'negative' evidence, i.e. evidence that is either not statistically significant or that conflicts with the prevailing beliefs about association under investigation. Under-representation occurs when researchers do not submit their study results to a journal (selective reporting), journals decide not to publish certain studies (selective publication), or when the study retrieval and selection procedures of a meta-analysis do not include a publication (selective inclusion)[6]. The bias caused by these phenomena is often referred to as publication bias although it is literally a 'pars pro toto' and dissemination bias may be more accurate[7].

Graphical assessments of heterogeneity and publication bias are numerous (as we will explain below). Some plots, such as the funnel plot and L'Abbé plot, have been evaluated, to some extent, individually[1,3,8-17]. However, little is known about the relative performance of the majority of graphs in terms of reproducibility (inter-rater agreement) and validity of the judgments of whether or not there is evidence of heterogeneity and publication bias. Our objective was to conduct a comprehensive evaluation of the reproducibility and validity of such judgments

with simulated data sets (based on empirical data) with varying heterogeneity and publication bias.

4.2 Methods

4.2.1 Frequently used graphs in meta-analysis

Figure 4.1 shows the most commonly used graphs in meta-analyses. Each graph is constructed from the same meta-analytical data set, i.e. the meta-analysis by Colditz et al. on the efficacy of the BCG vaccine in preventing tuberculosis[18].

The 'forest plot' (Figure 4.1a) is the most common graph in reports of meta-analyses. It shows the estimate (often a risk or odds ratio) and the confidence interval per study, and commonly the corresponding summary estimate. The plot dates back to at least the 1970s[19] and was used for the first time in a meta-analysis context in 1982[20]. In early use of the plot, studies and their confidence intervals were indicated by bars and studies with the largest variance (and thus largest confidence intervals and bars) were most prominent, even though they were least important. This problem was solved by displaying the confidence intervals with thin lines and marking a study's effect estimate with a square that was proportional to the study's weight (and inversely proportional to the variance). The summary or meta-analysis estimate evolved into a diamond, with its center at the summary estimate and its outer edges at the confidence limits. Forest plots may be useful for showing how the effect estimates from individual studies accumulate to a meta-analysis result. It has also been suggested that the plots provide a visual representation of the amount of variation between study estimates[3,21], and that eye-balling the overlap of the confidence intervals of the effect estimates can be used to judge the presence of between-study heterogeneity[22].

The 'funnel plot' was introduced by Light and Pillemer in 1985[4]. It typically has a measure for effect size on the x-axis and a measure related to the within-study variance (e.g. inverse standard error) on the y-axis. Each study is represented by a single equally-sized dot (Figure 4.1b). Under most circumstances, there are relatively more small studies (with larger variance) than big studies in a meta-analytical data set and the smaller studies have estimates that are more scattered and further removed from the summary estimate. This creates a funnel-like distribution of the dots in the plot that, without publication bias, is assumed to be symmetrical. If studies are relatively more 'missing' on one side of the plot - commonly studies with low sample size - the missingness and subsequent asymmetry can be attributed to publication bias. Because funnel plots are

constructed in many ways (with different measures on the axes) and asymmetry can be caused by more than just publication bias[23], it has been subject of substantial methodological scrutiny. Tang and Liu[15] assessed various funnel plots from published meta-analyses and concluded that the choice of axis measure could alter conclusions on presence of publication bias in most studies. Sterne et al.[14] evaluated the performance of funnel plots for binary event meta-analyses and concluded that the (inverse) standard error is the preferred measure for the y-axis and a log-transformed ratio measure of effect size (e.g. log risk ratio or log odds ratio) for the x-axis, as shown in Figure 4.1b. Finally, interpretations of funnel plots may be different for different readers[10,16].

The 'trim-and-fill plot' (Figure 4.1c) is also a funnel plot, although it involves not merely a representation of raw data but also shows the results of a kind of imputation algorithm called 'trim-and-fill'[24,25]. The algorithm assesses the symmetry of the plot analytically and can impute new studies to plots in which studies appear to be missing. The asymmetry assessment is performed via a rank correlation, after which the studies causing the asymmetry are trimmed on one side of the plot. This shifts the meta-analysis estimate, possibly causing asymmetry again. If re-estimation shows residual asymmetry, the process of trimming and re-estimation is repeated and usually runs for three to four cycles. Once no asymmetry is left, the trimmed studies are put back and their counterparts on the other side of the last symmetry axis are imputed. This is followed by a meta-analysis that includes the imputed studies. The plot itself looks like a regular funnel plot, but due to the imputation algorithm it contains additional dots (usually not filled) and an additional vertical line that indicates the summary effect when the imputed studies are included in the meta-analysis. The number of imputed studies and in particular the difference between the original summary estimate and the summary estimate after imputation are assumed indicative of publication bias. Although the algorithm's performance decreases when heterogeneity increases[25-27], in the presence of publication bias meta-analyses with imputations from the trim-and-fill algorithm are less biased than meta-analyses without imputations[26].

The 'Galbraith plot' (Figure 4.1d) is designed to assess the extent of heterogeneity between studies in a meta-analysis[1]. The y-axis shows the (log-transformed) effect size divided by its standard error (*z* score) and the inverse of the standard error on the x-axis. Each study is represented by a single dot and a regression line runs centrally through the plot. Parallel to the regression line, at two standard deviations distance, two lines create an interval in which most dots are expected to fall if the studies were to be estimating a single fixed parameter. Galbraith originally proposed to put the y-axis on the right side and make it radial so the

graph looks like a speedometer; the so-called radial plot. This is not necessary for interpretation of the graph and only few meta-analysis programs have implemented this feature[28].

The 'L'Abbé plot' (Figure 4.1e), introduced in 1987[2], is only applicable to meta-analyses of studies with binary outcomes. It plots the risks (or odds) in the exposed or index group (y-axis) against those of the control group (x-axis) and often contains a regression line and a central diagonal line indicating identical risks in each group. The size of the dots is proportional to the study weights. With a lot of between-study heterogeneity there may be substantial spread around the regression line, although it has been reported that naive use of the distance between the regression line and the dots as indicator of heterogeneity may be misleading[11]. Multiple clustering of control group risks along the x-axis indicates violation of the assumption that a single underlying baseline risk exists for all studies and thus also indicates heterogeneity.

In statistics, 'normal quantile plots' are often used to assess data normality. The normal quantile plot for meta-analysis, as recommended by Wang and Bushman[17], has each individual study's *z* score on the y-axis and the normalized quantile of its rank on the x-axis (Figure 4.1f). The plot can be used in meta-analysis to check normality of the data (dots expected on a straight line), to investigate heterogeneity (clustering of dots), and to assess the presence of publication bias (deviation of the tails from the regression line)[17].

The 'box plot' is, like the normal quantile plot, also frequently used in statistics[29]. Application to meta-analytical data sets may nevertheless be suboptimal because it does not take into account the weight of the studies. Consequently, the center (median or mean) does not correspond to the meta-analysis summary estimate and is therefore misleading. There are two adjusted box plots that do not have these disadvantages: (1) a standard box plot to which the meta-analysis summary measure has been added (not shown) and (2) a box plot of weighted estimates (Figure 4.1g)[30].

The 'standardized residual histogram' (Figure 4.1h) is briefly explained in a meta-analysis context by Greenland[31] and further described by Sutton et al.[32]. The histogram plots the fractions of categorized standardized residuals (the individual estimate minus the summary estimate, divided by the standard error of the individual estimate) in vertical bars. An overlay of a normal distribution can then be used to assess heterogeneity and departures from normality.

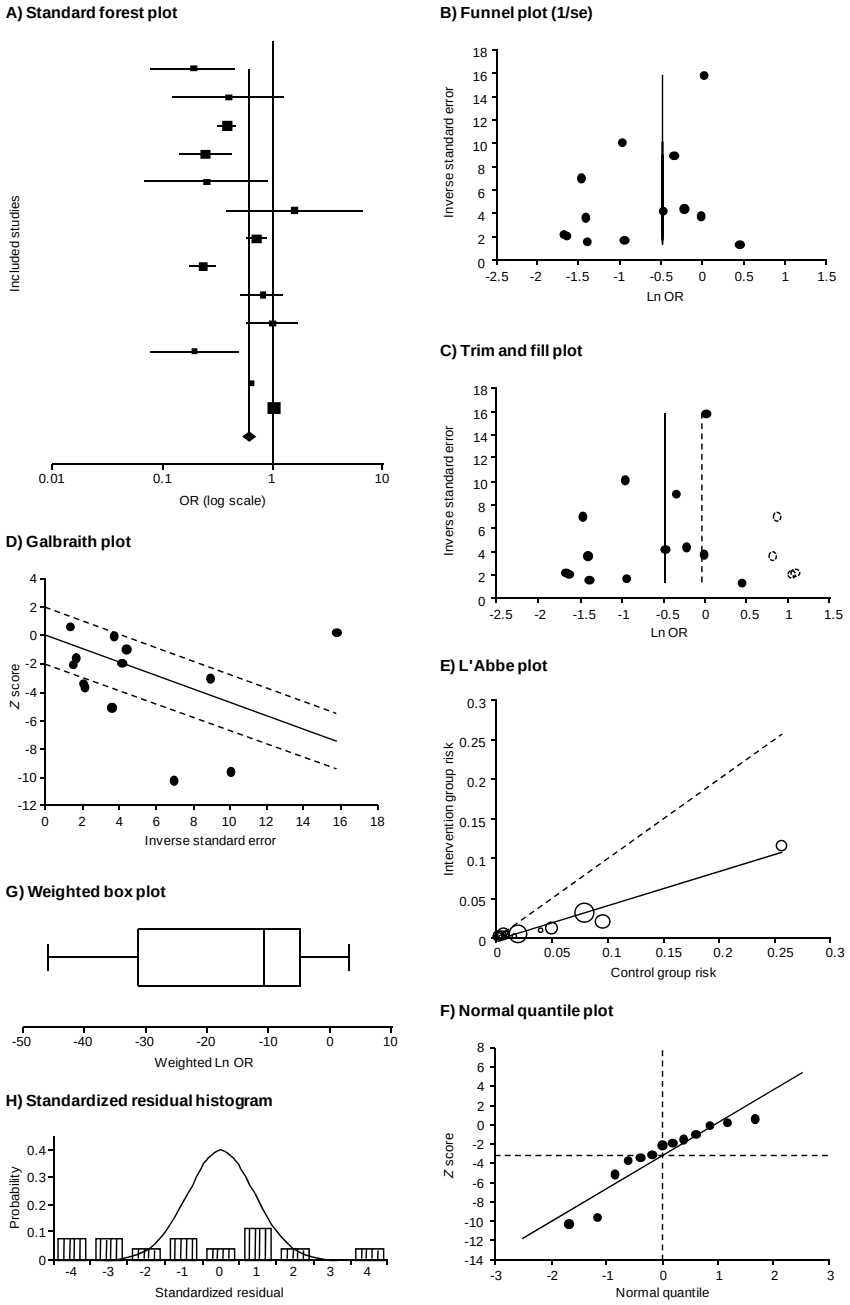


Figure 4.1 Meta-analysis graphs

The eight graphs correspond to the plots that are available in meta-analysis software for the assessment of heterogeneity or publication bias in meta-analysis. Two variations of the funnel plot and one variation of the box plot have been included in our assessments but are not shown here.

Other plots worth mentioning are the Egger regression asymmetry plot[9], the exclusion sensitivity plot[30] and the Baujat plot[8]. Egger's plot is essentially a Galbraith plot with the intercept of the regression line unconstrained and was therefore not added to the line-up of plots included in our assessments. The other two plots assess the influence of (excluding) a study on the total analysis; although useful for checking which studies are causing most of the heterogeneity and the largest shifts in summary estimates, they are not meant to quantify the heterogeneity itself and are also not further addressed in this study.

4.2.2 Reproducibility and validity of meta-analytical graphs

Assessment of graphs is inherently a subjective task and thus suffers from reproducibility issues. Reproducibility refers to the extent that a procedure, measurement or judgment is replicated either over time by the same persons or equipment or by different persons or under different circumstances. Here, reproducibility of the judgment of meta-analytical graphs refers to different judgments by different observers or raters (inter-observer or inter-rater agreement). Validity, in this context, refers to how well a graph measures what it is supposed to measure. It reflects both how well the judgment of a graph is guided by changes in heterogeneity and publication bias and the ability of the raters to properly interpret what is shown in the graph.

4.2.3 Simulation study

We designed a simulation study in which we created 100 meta-analysis data sets for each of four scenarios with different heterogeneity and publication bias (Table 4.1). The simulation parameters were based on those used in previous meta-analytical research[26,33,34] and on data from the 50 most recent (July 2007) English-language meta-analyses of therapeutic studies reporting an odds ratio in PubMed. Publication bias was induced by censoring studies with a commonly used exponential selection function[26,34,35] that gives studies with a high *P* value a higher chance to be censored.

We used the MIX meta-analysis software[30,36], incorporating a Visual Basic for Applications (VBA) version of the Mersenne Twister random number generator[37], to produce the data sets with randomly varying the characteristics described in Table 4.1. Scenarios 1 and 2 produced data sets without and scenarios 3 and 4 produced data sets with publication bias. Data sets from scenarios 1 and 3 exhibited little or no statistical between-study variation (heterogeneity) and data sets from scenario 2 and 4 produced data sets with substantial heterogeneity. For each simulated data set we created the eight graphs displayed in Figure 4.1, plus two additional funnel plots (with the standard error or sample size on the y-axis),

and an additional box plot, described above. The 11 graphs were created for each of the 100 simulated data sets, resulting in 1100 graphs that needed to be assessed. Numerical analyses, such as Mantel-Haenszel meta-analyses of odds ratios, were performed simultaneously.

Table 4.1 Simulation parameters

Parameter	Parameter distribution	Scenario 1 (H- / PB-)	Scenario 2 (H+ / PB-)	Scenario 3 (H- / PB+)	Scenario 4 (H+ / PB+)
Baseline risk	Uniform (min, max)	min=0.2 max=0.4	min=0.2 max=0.4	min=0.2 max=0.4	min=0.2 max=0.4
Odds ratio	LogNormal (mu, tau)	exp(mu)=0.85 tau=0	exp(mu)=0.85 tau=0.45	exp(mu)=0.85 tau=0	exp(mu)=0.85 tau=0.45
Number of studies	Uniform (min, max)	min=5 max=25	min=5 max=25	min=5 max=25	min=5 max=25
Study arm size	Uniform (min, max)	min=25 max=500	min=25 max=500	min=25 max=500	min=25 max=500
Study selection	Bernoulli (ps) = $\exp(-1b*p_i^a)$	a=0 b=0	a=0 b=0	a=3 b=4	a=3 b=4

H- = heterogeneity absent; PB- = publication bias absent; H+ = heterogeneity present, PB+ = publication bias present.

4.2.4 Graphical assessments

A special VBA program (available upon request) was written in which all graphs were presented to three researchers in random order and in a blinded fashion, meaning that the researchers were kept unaware of the source data (meta-analytical studies) of the graphs. Three raters (LB, NF, YY), with extensive experience in meta-analysis, scored the heterogeneity and publication bias shown by the graphs from 'None' to 'Extensive' with a continuous rating instrument (a scrollbar sliding from 0 to 100) inside the program. The ratings were performed over a period of three weeks.

4.2.5 Data analysis

We had two primary outcomes of interest in our study: the inter-rater reproducibility and the validity of the graphical assessments to judge the presence of heterogeneity and publication bias. Reproducibility was evaluated by means of intra-class correlation coefficients (ICCs) in SPSS 14.0[38]. The ICC can be viewed as a measure of correlation, consistency or conformity for a data set when it has multiple groups[39] and we used a two-way random effects ICC for consistency of individual measurements[39]. ICCs range from 0 to 1, typically with the following classification: ICC < 0.75 = poor agreement, 0.75 < ICC < 0.90 = moderate agreement, and ICC > 0.90 = high agreement[40].

The validity of the graphical judgments of heterogeneity and publications bias was evaluated with regression analyses. The dependent variables were the presence or absence of between-study variability (heterogeneity) and the counts of censored studies (publication bias), both determined by the simulation settings. The average score of the three raters was used as independent (predictor) variable and the logistic and Poisson regression analyses were performed in R[41].

4.3 Results

4.3.1 Reproducibility

The overall ICC (across all raters and all 1100 graphs) was 0.58 (95% confidence interval (CI): 0.54, 0.62) for assessments of heterogeneity and 0.65 (95% CI: 0.62, 0.69) for assessments of publication bias. With regard to heterogeneity, the forest plot and the standardized residual histogram had the highest reproducibility (ICCs of 0.87 and 0.69, respectively). The weighted box plot ratings were least reproducible. For judgment of publication bias, the trim-and-fill plot and the weighted box plot had the best reproducibility (ICCs of 0.91 and 0.71) and the standard error funnel plot and the normal quantile plot had the lowest ICCs. Details are provided in Table 4.2.

Table 4.2 Reproducibility

	Heterogeneity ICC[a]		**Publication bias ICC[a]**	
Graph	**Estimate**	**CI**	**Estimate**	**CI**
box plot	0.61	0.506, 0.703	0.59	0.488, 0.689
weighted box plot	0.34	0.211, 0.462	0.71	0.622, 0.783
residual histogram	0.69	0.600, 0.768	-	
normal quantile plot	0.47	0.353, 0.585	0.50	0.387, 0.613
forest plot	0.87	0.823, 0.905	-	
SE funnel plot	-	-	0.51	0.397, 0.621
inverse SE funnel plot	-	-	0.53	0.412, 0.632
sample size funnel plot	-	-	0.58	0.473, 0.679
Galbraith plot	0.63	0.527, 0.718	-	-
L'Abbé plot	0.55	0.436, 0.651	-	-
trim-and-fill plot	-	-	0.91	0.874, 0.934
overall	0.58	0.544, 0.621	0.65	0.617, 0.686

CI = confidence interval, SE= standard error

[a] Two-way, random effects intraclass correlation coefficients.

4.3.2 Validity

For assessment of heterogeneity, the scores of the forest plot, standardized residual histogram, Galbraith plot, and L'Abbé plot showed significant association with the presence of heterogeneity (Table 4.3). When taking their ICCs also in

consideration, the standardized residual histogram and the forest plot would be the best candidates if multiple graphs were to be used. For the assessment of publication bias, validity was low in general. None of the funnel plots correlated well with the underlying parameters of publication bias.

Table 4.3 Validity

Graph	Regression coefficient[a]	Coefficient *P* value
Heterogeneity (logistic regression)		
box plot	0.009	0.29
weighted box plot	-0.015	0.22
standardized residual histogram	0.115	<0.001
normal quantile plot	0.001	0.9
forest plot	0.096	<0.001
Galbraith plot	0.094	<0.001
L'Abbé plot	0.047	<0.001
Publication bias (Poisson regression)		
box plot	0	0.96
weighted box plot	-0.006	0.10
normal quantile plot	-0.001	0.74
standard error funnel plot	0.001	0.69
inverse standard error funnel plot	0.001	0.78
sample size funnel plot	0.001	0.71
trim-and-fill plot	0.001	0.83

[a] Univariable association of the graph scores (independent variables) and the simulation parameters for heterogeneity and publication bias (dependent variables).

4.4 Discussion

We examined the inter-rater reproducibility and validity of 11 graphs that are frequently used in assessments of heterogeneity and publication bias in meta-analysis. A hundred data sets with varying heterogeneity and publication bias were simulated and the resulting 1100 graphs were judged in random order by three raters on the degree of heterogeneity and/or publication bias. The reproducibility of heterogeneity assessments was highest for the forest plot and the standardized residual histogram. Association between graph ratings and publication bias (censorship of studies) was poor, with the weighted box plot and the trim-and-fill plot showing the best combination of reproducibility and validity. Reproducibility depended heavily on the type of graph and only a few graph ratings validly reflected underlying heterogeneity and publication bias.

The issue of poor reproducibility of graph assessments in meta-analyses has been raised by others[10,15,16], in particular for the funnel plot. However, it had not yet been formally investigated or quantified. Our results underline the need to use

multiple raters and come to composite or consensus scores. We acknowledge that our results are likely to be sensitive to the experience and training of the raters and to some extent to the number of raters. The reproducibility and validity of graph assessments will probably decrease if the reviewers or meta-analysts are less experienced and perhaps increase when they are all experts. We consequently decided to use raters that have an experience and a background in meta-analysis that is common for authors of systematic reviews (one experienced rater (LB) and two raters with moderate experience in meta-analysis (NF, YY)). We used three raters as this is a common number for systematic review teams. Both aspects should make our findings generalizable to the average practice of meta-analysis.

We found that the forest plot and the standardized residual histogram (plots with the best reproducibility) indicate the presence or absence of heterogeneity well. Although eye-balling of the forest plot is traditionally not recommended for investigating heterogeneity[5], we found in an additional explorative analysis (data available upon request) that ratings of the forest plot as well as the histogram correlated very strongly with the results of the I^2 test[42,43], which is commonly used to quantify heterogeneity in meta-analyses.

To simulate publication bias, we used a common selection approach based on an exponential function of the *P* value[26,34,35]. This assumes that censoring is related to (absence of) significance, whereas the funnel plots assume censorship based on (absence of) extremeness of small-study estimates. These premises may in many situations assume censoring of different studies, which could explain the relatively poor performance of these plots in our validity evaluation. Although our approach is common and likely to be close to reality, future simulation studies might explore alternative censorship mechanisms, including possible differences in these mechanisms for experimental and observational studies.

To put the analytics of this paper in perspective, we would like to stress that prospective registries of studies are essential to gain knowledge on why certain studies are published and others not. Tracking of study reporting via these registries of protocols is the only way to prevent or properly correct for publication biases. Time will tell whether this is a utopia or a future reality.

There is a well-known expression that says “Pictures say more than a thousand words”. We would like to add that, in meta-analysis, they may say more than a million numbers. Interpretation of the graphs requires care, however, because reproducibility and validity depend heavily on the type of graph. The forest plot and standardized residual histogram are the best candidates to visualize

heterogeneity, with the trim-and-fill plot and the weighted box plot as their counterparts in explorations of publication bias.

Acknowledgments

This study was supported by research grant 3084 from the Graduate School of Medical Sciences of Kitasato University. Kitasato University played no role in any aspect of the study.

Conflict of interest

None declared.

References

1. Galbraith RF. A note on graphical presentation of estimated odds ratios from several clinical trials. *Stat Med.* Aug 1988;7(8):889-894.

2. L'Abbe KA, Detsky AS, O'Rourke K. Meta-analysis in clinical research. *Ann Int Med.* Aug 1987;107(2):224-233.

3. Lewis S, Clarke M. Forest plots: trying to see the wood and the trees. *BMJ.* Jun 16 2001;322(7300):1479-1480.

4. Light RJ, Pillemer DB. *Summing up*. Cambridge: Harvard University Press; 1985.

5. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ.* Nov 19 1994;309(6965):1351-1355.

6. Rothstein H, Sutton AJ, Borenstein M. *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. 1st ed. Chichester: Wiley; 2005.

7. Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ. Publication and related biases. *Health Technol Assess.* 2000;4(10):1-115.

8. Baujat B, Mahe C, Pignon JP, Hill C. A graphical method for exploring heterogeneity in meta-analyses: application to a meta-analysis of 65 trials. *Stat Med.* Sep 30 2002;21(18):2641-2652.

9. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ.* Sep 13 1997;315(7109):629-634.

10. Lau J, Ioannidis JP, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. *BMJ.* Sep 16 2006;333(7568):597-600.

11. Song F. Exploring heterogeneity in meta-analysis: is the L'Abbe plot useful? *J Clin Epidemiol.* Aug 1999;52(8):725-730.

12. Song F, Khan KS, Dinnes J, Sutton AJ. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. *Int J Epidemiol.* Feb 2002;31(1):88-95.

13. Souza JP, Pileggi C, Cecatti JG. Assessment of funnel plot asymmetry and publication bias in reproductive health meta-analyses: an analytic survey. *Reprod Health.* 2007;4:3.

14. Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *J Clin Epidemiol.* Oct 2001;54(10):1046-1055.

15. Tang JL, Liu JL. Misleading funnel plot for detection of bias in meta-analysis. *J Clin Epidemiol.* May 2000;53(5):477-484.

16. Terrin N, Schmid CH, Lau J. In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *J Clin Epidemiol.* Sep 2005;58(9):894-901.

17. Wang MC, Bushman BJ. Using the normal quantile plot to explore meta-analytic data sets. *Psychology Methods.* 1998;3(1):46-54.

18. Colditz GA, Brewer TF, Berkey CS, et al. Efficacy of BCG vaccine in the prevention of tuberculosis. Meta-analysis of the published literature. *JAMA.* Mar 2 1994;271(9):698-702.

19. Freiman JA, Chalmers TC, Smith H, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: survey of 71 "negative trials". *N Engl J Med.* 1978;299:690-694.

20. Lewis JA, Ellis SH. A statistical appraisal of postinfarction betablocker trials. *Primary Cardiol.* 1982(supp1):31-37.

21. Moja L, Moschetti I, Liberati A, Gensini GF, Gusinu R. Understanding systematic reviews: the meta-analysis graph (also called 'forest plot'). *Intern Emerg Med.* 2007;2(2):140-142.

22. Ried K. Interpreting and understanding meta-analysis graphs--a practical guide. *Aust Fam Physician.* Aug 2006;35(8):635-638.

23. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *J Clin Epidemiol.* Nov 2000;53(11):1119-1129.

24. Duval SJ, Tweedie RL. A non-parametric 'trim and fill' method of accounting for publication bias in meta-analysis. *J Am Stat Assoc.* 2000;95:89-98.

25. Duval SJ, Tweedie RL. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics.* Jun 2000;56(2):455-463.

26. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Stat Med.* May 2 2007.

27. Sterne JA. High false-positive rate for trim and fill method. *Electronic response to Sutton et al. in BMJ 2000;320:1574-1577.* 2000:*http://www.bmj.com/cgi/eletters/320/7249/1574#EL7241*.

28. Bax L, Yu LM, Ikeda N, Moons KG. A systematic comparison of software dedicated to meta-analysis of causal studies. *BMC Med Res Methodol.* 2007;7(1):40.

29. Tukey JW. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley; 1977.

30. Bax L, Yu LM, Ikeda N, Tsuruta H, Moons KG. Development and validation of MIX: comprehensive free software for meta-analysis of causal research data. *BMC Med Res Methodol.* 2006;6:50.

31. Greenland S. Quantitative methods in the review of epidemiologic literature. *Epidemiol Rev.* 1987;9:1-30.

32. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. *Methods for Meta-Analysis in Medical Research*. Chichester: Wiley; 2000.

33. Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Stat Med.* Oct 30 2006;25(20):3443-3457.

34. Macaskill P, Walter SD, Irwig L. A comparison of methods to detect publication bias in meta-analysis. *Stat Med.* Feb 28 2001;20(4):641-654.

35. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics.* Dec 1994;50(4):1088-1101.

36. Bax L, Yu LM, Ikeda N, Tsuruta H, Moons KGM. MIX: comprehensive free software for meta-analysis of causal research data. Version 1.7. *http://mix-for-meta-analysis.info*.

37. Matsumoto M, Nishimura T. Mersenne Twister, a very fast random number generator. *http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/emt.html*.

38. *SPSS for Windows* [computer program]. Version 14.0. Chicago, USA: SPSS Inc.; 2005.

39. McGraw KO, Wond SP. Forming inferences about some intraclass correlation coefficients. *Psychology Methods.* 1996;1(1):30-46.

40. Portney LG, Watkins MP. *Foundations of Clinical Research. Applications to Practice.* 2nd ed. Upper Saddle River, NJ: Prentice Hall Health; 2000.

41. *R: A language and environment for statistical computing* [computer program]. Version. Vienna, Austria; 2007.

42. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med.* Jun 15 2002;21(11):1539-1558.

43. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ.* Sep 6 2003;327(7414):557-560.

5

DEVELOPMENT OF SOFTWARE FOR CAUSAL META-ANALYSIS

Based on:

Bax L, Yu LM, Ikeda N, Tsuruta H, Moons KG.
Development and validation of MIX: comprehensive free software for meta-analysis of causal research data.
BMC Med Res Methodol 2006, 6(50).

Abstract

Introduction. Meta-analysis has become a well-known method for synthesis of quantitative data from previously conducted research in applied health sciences. So far, meta-analysis has been particularly useful in evaluating and comparing therapies and in assessing causes of disease. Consequently, the number of software packages that can perform meta-analysis has increased over the years. Unfortunately, it can take a substantial amount of time to get acquainted with some of these programs and most contain little or no interactive educational material.

Methods. We set out to create and validate an easy-to-use and comprehensive meta-analysis package that would be simple enough programming-wise to remain available as a free download. We specifically aimed at students and researchers who are new to meta-analysis, with important parts of the development oriented towards creating internal interactive tutoring tools and designing features that would facilitate usage of the software as a companion to existing books on meta-analysis.

Results. We took an unconventional approach and created a program that uses Excel as a calculation and programming platform. The main programming language was Visual Basic, as implemented in Visual Basic 6 and Visual Basic for Applications in Excel 2000 and higher. The development took approximately two years and resulted in the 'MIX' program, which can be downloaded from the program's website free of charge. Next, we set out to validate the MIX output with two major software packages as reference standards, namely STATA (metan, metabias, and metatrim) and Comprehensive Meta-Analysis Version 2. Eight meta-analyses that had been published in major journals were used as data sources. All numerical and graphical results from analyses with MIX were identical to their counterparts in STATA and CMA. The MIX program distinguishes itself from most other programs by the extensive graphical output, the click-and-go (Excel) interface, and the educational features.

Conclusion. The MIX program is a valid tool for performing meta-analysis and may be particularly useful in educational environments. It can be downloaded free of charge via http://www.mix-for-meta-analysis.info.

5.1 Introduction

The amount of data produced by researchers in health sciences has been growing explosively and advances in genetics, genomics, and information technology are likely to further contribute to this growth. In the past two decades, meta-analysis has evolved into the statistical method par excellence to make sense out of the growing number of research reports. As the quantitative analytical part of a systematic review, it has been used for evaluating data from both experimental and observational studies in therapeutic, diagnostic, prognostic, and etiologic settings. In the commonly used definition of the hierarchy of scientific data for medical decision making, meta-analyses are considered as providing the highest level of evidence[1,2]. As such, they can have a major impact on medical practice and health care policies, especially if aggregating data and investigating sources of heterogeneity provide new insights. Two well-known examples are the meta-analyses by Yusuf et al.[3] and Lau et al.[4], both showing that meta-analysis can be a powerful tool to show intervention effects that would remain beneath the surface of single study data without proper synthesis and re-analysis.

Although meta-analyses can be applied to all types of medical research, its primary application so far has been in the therapeutic realm. One of the main forces behind the rise of therapeutic meta-analysis is the Cochrane Collaboration[5], whose effort to systematically assess and synthesize evidence from randomized controlled trials has so far produced more than 4400 Cochrane systematic reviews, many with quantitative meta-analyses. The increasing interest for meta-analysis in health sciences over the past twenty years has been reported by several authors[6-11] and a small search we did in preparation of this project reveals that between 1990 and 2005 approximately 12,000 publications have been classified as a meta-analysis by PubMed.

Many general statistical software packages have included options for meta-analysis in their basic program configuration, and user-communities have written numerous meta-analysis add-ons. Specialized software packages, meant exclusively for meta-analysis, are also available in various types and price ranges. Although the number of software packages for performing meta-analysis is substantial, in our opinion, most share one common limitation: low applicability in educational settings or environments with beginning researchers. Even though numerous researchers in health care are nowadays confronted with data from published meta-analyses or are even requested to do a meta-analysis themselves, there is still little or no electronic educational material and none of the existing software has explicit educational features. Cost is another issue that may have an

impact on the use of software by students and lecturers: only a few of the modern meta-analysis packages are free and if academic pricing is available, prices can still be rather high for many.

After reading previously published software reviews[12-15] and using existing meta-analysis software, we made an inventory of what we thought was lacking or could be improved. Next, we set out to implement our ideas and create an innovative and comprehensive statistical meta-analysis package that would be freely accessible and user-friendly enough for students and beginning researchers. The program, called MIX (Meta-analysis with Interactive eXplanations), has been developed over the past two years and has been presented at several stages of the development at a number of conferences[16-19]. In October 2005, the first public version (1.0) was released during the Cochrane Colloquium in Melbourne[18] and has become available for download via the MIX website[20]. It has been receiving a lot of interest (100–150 unique visitors to the MIX website each week) and has been downloaded over 1800 times within 6 months of its first release. This has prompted us to validate the results of all tests in the program formally and this article provides the official introduction of the MIX program together with the results of the validation.

5.2 Methods

5.2.1 Objectives

Our primary objective was to develop a free program for meta-analysis of causal research (therapeutic trials as well as etiologic cohorts and case-control studies) that could be applied in both analytical and educational settings. Our secondary aim was to validate the analytical tests in the program with output from established reference standards.

5.2.2 Program development

Before the actual development, we started with making an inventory of the most important meta-analytical tests and approaches, and brainstormed on ideas for an interface. Since causal meta-analysis methods are relatively well-established (in contrast to diagnostic or prognostic approaches to meta-analysis), we focused on meta-analysis of controlled trials and cohort or case-control studies. In these studies, outcome differences between exposed or treated and non-exposed or untreated groups are compared to assess a causal relationship between the determinant (treatment or exposure) and an outcome (mortality or morbidity). As far as the program structure was concerned, our a priori idea was to create an add-in for Excel. Although a rather unorthodox approach in this area (all existing meta-analysis programs are stand-alone programs and work independently of

Microsoft Office), Excel provides a sophisticated calculation and graphics platform that is well-suited to many meta-analytical methods and at the programmer's disposal before any programming is done. Consequently, development and maintenance is relatively easy and costs can be kept to a minimum (one of the main aims in our program development). Furthermore, the spreadsheet environment of Microsoft Excel is familiar to almost all researchers in medical, social, and economical sciences, which was very much in line with our attempt to develop a package that is fit for beginning researchers. Although we realized that even recent versions of Excel can be inaccurate with regard to some statistical calculations[21-23], we were confident that we could program around these difficulties if necessary.

Since we wanted to move beyond the occasional spreadsheet that can perform meta-analytical calculations, we started by designing a programming structure in which the already existing Excel functionality could be exploited to its maximum. Sophisticated procedures were custom-programmed with Visual Basic in the Visual Basic for Applications (VBA) editor of Excel 2003 (and tested in Excel 2000 and onward). The so-called front-loader (a start-up program initiated with an icon) and some small assistant programs, all being non-Excel entities, were developed with Visual Basic 6.0 (VB6).

5.2.3 Program architecture and operation

The current version of the program (version 1.5) is still only compatible with Windows operating systems running Excel 2000 or later, but versions for use with Excel on Macintosh and Linux are in preparation. The descriptions below apply to the Windows version, though most of it can be extended to future versions for other operating systems.

Installation is made easy with a set-up program that installs the necessary files in a folder that can be specified by the user (default is C:\Program Files\MIX). It will also create a MIX item in the Windows Start Menu (installing additional start-up icons on the Desktop or in the Quick-Launch bar is optional) and provides the option to start a Flash®-based program introduction. The MIX menu item contains an icon for starting up the MIX program, a folder with a shortcut to the uninstall program, a folder with shortcuts to programs for loading and unloading the Excel add-in, and a folder with educational programs and information. Loading the small MIX add-in that is supplied with the main program (typically automatically loaded during installation) results in a MIX menu-item under the Tools menu in Excel. This MIX menu contains several functions that can be accessed when the MIX program itself is not running. The files that form the core of the program are

recognizable by their MIX file extension (*.mix) and currently contain approximately 16,000 lines of command code in 26 code modules and 17 custom user forms. These core files take up approximately 22 Mb of space on a hard-disk and their primary functions are (A) running interface procedures, (B) showing and manipulating output, (C) performing analyses, and finally (D) exporting and communicating with external files and programs. One of the core files is a large Excel workbook with 23 worksheets that forms the calculation engine of the program. It contains six sheets with primarily worksheet formulas and ten sheets with various kinds of pre-calculated graphical and numerical results from meta-analytical tests. The remaining sheets contain information for help functions or programming purposes. This Excel workbook remains hidden from the users at all times.

At start-up, a dedicated instance (an independent fully functional running program) of Excel is created and becomes visible once all regular Excel menus and toolbars are hidden and replaced by the MIX graphical interface. The Excel instance used by MIX is secured for exclusive use by the MIX program and does not interfere with existing Excel windows or settings.

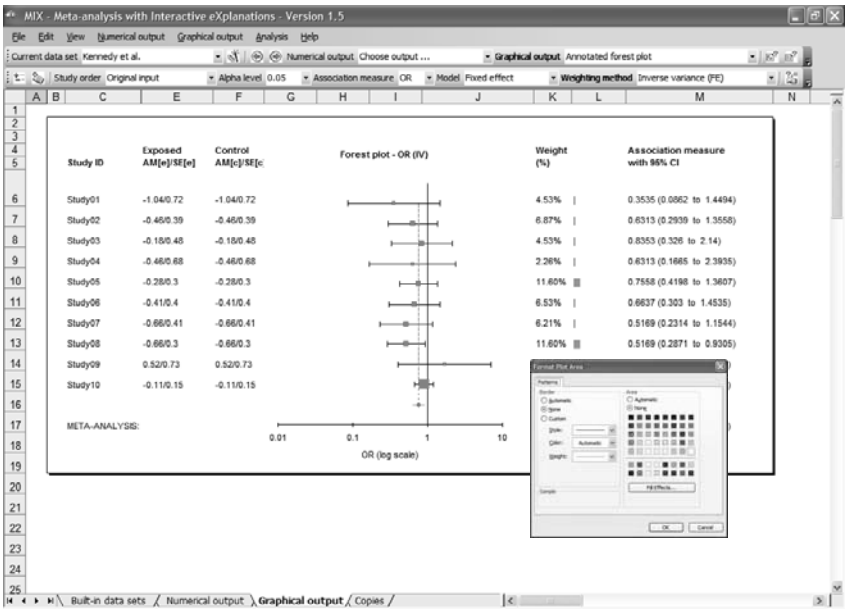


Figure 5.1 MIX user interface with an annotated forest plot

The standard Excel menu and toolbars have been replaced by the MIX interface. MIX can produce a wide variety of graphs and numerical output. The forest plot above is annotated with per arm data, weights of the studies in the analysis (including a weight bar), the per study association measures, and at the bottom right the meta-analysis estimate. Right-clicks and double clicking graphical items shows the formatting options that Excel users are familiar with.

The interface consists of a menu bar, two toolbars, and several shortcut menus. The menu bar and toolbar are directly accessible and the shortcut menus pop up with a right click of the mouse. The MIX menu bar has eight main menus (File, Edit, View, Numerical Output, Graphical Output, Analysis, and Help) via which all functions of the MIX program can be executed. Most of the common functions require only a single click on the toolbars. Double clicking graph items skips the shortcut menu and directly provides options for changing the graph item's format. Figure 5.1 shows the MIX program's user-interface with a forest plot.

The MIX program provides several options for importing or creating data sets for meta-analysis. The most convenient option is to create an Excel or CSV file with data (standard output option in Excel) and import this file into the MIX program. The variable ranges are then selected in Excel-manner to create a data set (see Figure 5.2), which is subsequently loaded for analysis and optionally saved as a MIX data set file (*.mxd). The program accepts descriptive data from studies with continuous outcomes, e.g. sample size, mean, standard deviation, and dichotomous outcomes, e.g. group sizes and event numbers (two-by-two table data). Comparative data can also be loaded by means of association measures with their standard error.

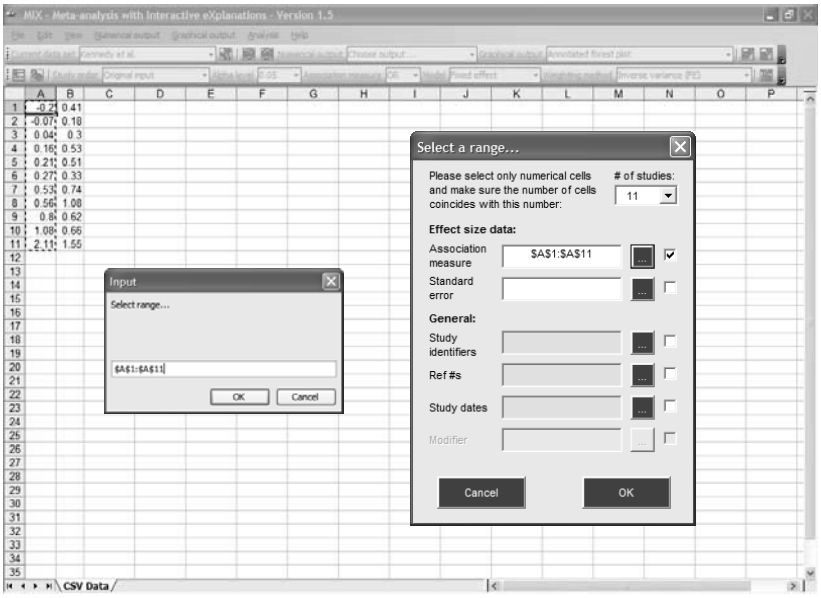


Figure 5.2 Creation of a data set in MIX via selection of ranges

There are three ways to enter data in MIX: 1) manually in an empty Excel sheet opened in MIX, 2) copied from another spreadsheet and pasted in an empty Excel sheet opened in MIX, and 3) imported from an existing data set in a CSV or Excel file. In all cases, dialog boxes similar to the ones floating over the Excel sheet shown above will pop up and allow the user to select the ranges corresponding to the statistics for each study that are to be entered in the meta-analysis.

Initially, however, it is not necessary to make a data set since 19 data sets from the most authoritative books on the subject ("Meta-analysis in Medical Research" by Sutton et al[10], "Systematic Reviews in Health Care, Meta-Analysis in Context" by Egger et al[6], and "Systematic Reviews in Health Care, A Practial Guide" by Glasziou et al[7]) have been included in the program. Most analyses and graphs presented in these books can be reproduced with a few clicks and the program can be used as learning or teaching companion to these books. We hope to support more books in this way in the future. In addition, the MIX website also contains a data set repository where users can contribute and download MIX data sets.

A large variety of numerical and graphical output can be produced by the program. Besides the association measure values from the meta-analysis, several formal tests for heterogeneity, small-study effects (publication bias), single study influence, and cumulative trends are also available in MIX. The graphical output is particularly comprehensive, with no less than eighteen informative plots that can be formatted in detail.

Possible association measures from continuous outcome data input are mean difference (MD), Hedges' *g* (HG), and Cohen's *d* (CD), analyzed by inverse variance fixed or random effects models. Data from studies with dichotomous outcomes can be analyzed with a risk difference (RD), risk ratio (RR), or odds ratio (OR), weighted by inverse variance, Mantel-Haenszel, Peto (only odds ratio), or DerSimonian-Laird approaches. Analyses based on correlation coefficients or Fisher's *Z* are also possible, though only if the data are provided as comparative input, e.g. the association measures with their standard error. If correlations or effect sizes are not in this format, they can be transformed via the MIX Statistics Converter that comes with the program. Table 5.1 gives an overview of general features and methods in version 1.5 of the MIX program.

The most important educational features are the program's Output Tutor and Concept Tutor. Both are interactive dialog boxes that provide information about epidemiological and statistical concepts and tests. The Output Tutor changes with each analysis and always explains tests and results that are displayed or changed at the very moment. Additional teaching material includes a Flash®-based Theory Tour that explains the fundamentals of systematic reviews and meta-analyses and a Program Tour that shows the basics of how to use the program. The educational materials take up approximately 25 Mb and can also be downloaded separately.

To increase program stability and prevent users from accidentally altering the Visual Basic procedures, the source code cannot be accessed while the program is

running. Codes to unlock the VBA modules are provided by the first author upon request.

Table 5.1 General features of MIX

Program
- Program size: 22Mb
- Compatibility: Windows, Excel 2000 or later
- Installation: Automated set-up (executable)
Data type compatibility
- Descriptive dichotomous (2x2 table data)
- Descriptive continuous (n, m sd)
- Comparative (am, se)
Data input options
- Manual
- Excel and CSV-files
- Effect size and statistic conversions

Export options
- Output to clipboard
- Data set to Excel or CSV
- Ready-made reports to Excel
- Multi-resolution graphs to clipboard
Learning items
- Output tutor
- Concept tutor
- Built-in data sets
- Theory guide (Flash)*
- Program guide (Flash)*
- HTML help

(*) The theory and program guides are available via the Internet. The link can be accessed from inside MIX as well as from the MIX menu in the Windows Start Menu.

5.2.4 Validation

Version 9.2 of STATA[24], and more specifically version 1.81 of the metan program[25], version 1.2.4 of the metabias program[26], and version 1.0.5 of the metatrim program[27] were used as the general reference standards for most tests. Details on the development of these user-written programs themselves can be found in the STATA Technical Bulletins[25-27]. The meta-analysis software Comprehensive Meta-Analysis (CMA) version 2[28] was used for validation of the Fail-safe *N* output and to double check the results of the other tests. Two investigators performed the validation independently with the MIX program (version 1.5 running in Excel 2003) and the reference standard(s) by analyzing eight data sets from meta-analyses that have been published in major journals[4,29-35].

The data sets represent three of the most often used types of data for meta-analysis in health care research: 1) descriptive data for dichotomous outcomes, 2) descriptive data for continuous outcomes, and 3) comparative (association measure) data. For all three data types we chose a relatively small (less than 10 studies) and large data set (more than 20 studies) and we used two extra data sets in the 'descriptive dichotomous' category (one representing a meta-analysis of substantially heterogeneous studies and one with a rare event). The data sets are summarized in Table 5.2. The tests that were subject to the validation procedures are shown in Table 5.3. The items include individual study association measures, combined association measures, and several heterogeneity and small-study effect

assessments. Whenever applicable, *P* values and/or confidence intervals were also compared.

Table 5.2 Data sets used in the validation

#	Author(s)	Date	Study	Input type
1	Lau et al.[4]	1992	33	DD – large
2	Hodnett et al.[36]	2001	5	DD – small
3	Teo et al.[37]	1991	16	DD – publication bias
4	Crowley[38]	2000	17	DD – rare events
5	Lightowler et al.[39]	2003	5	DC – small
6	Wahlbeck et al.[40]	2000	11	DC – medium large
7	Pagliaro et al.[33]	1992	19	C – odds ratio
8	Law et al.[41]	1994	10	C – risk difference

The validation was done with eight data sets from meta-analysis that have been published in major peer-reviewed journals. The data sets were selected to represent a wide spectrum of potential input for meta-analysis. Abbreviations: "DD" = descriptive data for dichotomous outcomes (two-by-two table data), "DC" = descriptive data for continuous outcomes (means with their standard deviations and sample sizes), and "C" = comparative data (association measures with standards error or confidence intervals).

Table 5.3 Statistics assessed during the validation

Study data (per association measure/weighting)
- Association measure with 95% CI and/or *P* value
- Study contribution weights
Meta-analysis (per association measure/weighting)
- Association measure with 95% CI and *P* value
- Heterogeneity *Q* with 95% CI and/or *P* value
- Inconsistency I^2 with 95% CI and/or *P* value
- Fail-safe N with tolerance level
- Begg's rank correlation test with *z*-score and *P* value
- Egger's intercept with 95% CI and/or *P* value
- Macaskill's slope with 95% CI and/or *P* value
- Trim and fill studies with new association measure and 95% CI

Essentially all major numerical output that is produced by a comprehensive meta-analysis was assessed during the validation. The tests were repeated with all available (fixed effect and random effects) weighting models. Abbreviations: "CI" = confidence interval.

Results from the analyses of the eight data sets with MIX and the reference software were entered independently in identical custom-made spreadsheets. These spreadsheets were later compared in separate analysis sheets that used a cell-based formula to check for discrepancies of results up to 4 decimals.

5.3 Results

In summary, we have been able to achieve our objective of developing a comprehensive and yet free program for meta-analysis. The Excel platform,

although not without problems, has proved to be flexible enough to create an easy-to-use, and graphically and numerically comprehensive program.

Table 5.4. Individual study weighting validation with data set 1.

	Weighting method and weights (%)							
	IV (FE)		MH (FE)		PETO (FE)		IV+t^2 (RE)	
Studies	**MIX**	**STATA**	**MIX**	**STATA**	**MIX**	**STATA**	**MIX**	**STATA**
Fletcher	0.07%	0.07%	0.18%	0.18%	0.11%	0.11%	0.21%	0.21%
Dewar	0.21%	0.21%	0.27%	0.27%	0.22%	0.22%	0.59%	0.59%
Euro 1	0.74%	0.74%	0.53%	0.53%	0.74%	0.74%	1.99%	1.99%
Euro 2	3.38%	3.38%	3.68%	3.68%	3.39%	3.39%	7.18%	7.18%
Heikinhe	0.95%	0.95%	0.74%	0.74%	0.95%	0.95%	2.50%	2.50%
Italian	0.89%	0.89%	0.76%	0.76%	0.88%	0.88%	2.35%	2.35%
Aust 1	1.39%	1.39%	1.39%	1.39%	1.38%	1.38%	3.50%	3.50%
Frankfurt	0.80%	0.80%	1.18%	1.18%	0.90%	0.90%	2.14%	2.14%
NHLBI	0.21%	0.21%	0.12%	0.12%	0.24%	0.24%	0.60%	0.60%
Frank	0.29%	0.29%	0.26%	0.26%	0.29%	0.29%	0.82%	0.82%
Valere	0.42%	0.42%	0.35%	0.35%	0.42%	0.42%	1.17%	1.17%
Klein	0.07%	0.07%	0.04%	0.04%	0.10%	0.10%	0.21%	0.21%
UK-Col	1.85%	1.85%	1.67%	1.67%	1.81%	1.81%	4.46%	4.46%
Austrian	2.23%	2.23%	2.64%	2.64%	2.34%	2.34%	5.21%	5.21%
Aust 2	1.14%	1.14%	1.24%	1.24%	1.13%	1.13%	2.94%	2.94%
Lasierra	0.07%	0.07%	0.14%	0.14%	0.09%	0.09%	0.20%	0.20%
N Ger Col	2.35%	2.35%	1.85%	1.85%	2.33%	2.33%	5.44%	5.44%
Witchitz	0.22%	0.22%	0.22%	0.22%	0.22%	0.22%	0.64%	0.64%
Euro 3	1.05%	1.05%	1.24%	1.24%	1.09%	1.09%	2.73%	2.73%
ISAM	2.96%	2.96%	2.74%	2.74%	2.92%	2.92%	6.50%	6.50%
GISSI-1	33.05%	33.05%	31.84%	31.84%	32.65%	32.65%	21.00%	21.00%
Olson	0.07%	0.07%	0.10%	0.10%	0.08%	0.08%	0.20%	0.20%
Baroffio	0.05%	0.05%	0.30%	0.30%	0.15%	0.15%	0.14%	0.14%
Schreiber	0.08%	0.08%	0.13%	0.13%	0.10%	0.10%	0.22%	0.22%
Cribier	0.05%	0.05%	0.04%	0.04%	0.05%	0.05%	0.15%	0.15%
Sainsous	0.20%	0.20%	0.26%	0.26%	0.22%	0.22%	0.57%	0.57%
Durand	0.17%	0.17%	0.19%	0.19%	0.17%	0.17%	0.47%	0.47%
White	0.18%	0.18%	0.54%	0.54%	0.35%	0.35%	0.51%	0.51%
Bassand	0.25%	0.25%	0.30%	0.30%	0.27%	0.27%	0.71%	0.71%
Vlay	0.06%	0.06%	0.09%	0.09%	0.07%	0.07%	0.19%	0.19%
Kennedy	0.71%	0.71%	0.78%	0.78%	0.71%	0.71%	1.90%	1.90%
ISIS-2	43.68%	43.68%	43.92%	43.92%	43.47%	43.47%	22.18%	22.18%
Wisenb	0.14%	0.14%	0.28%	0.28%	0.16%	0.16%	0.40%	0.40%

In a meta-analysis, each study is given a weight that determines its influence on the overall result and this weight depends on the weighting method. Proper weighting is crucial to get correct results, so we validated all individual study weights for each data set and weighting method. The table shows the odds ratio weighting validation for data set 1. Abbreviations: "IV" = inverse variance weighting, "MH" = Mantel-Haenszel weighting, "PETO" = Peto weighting, and "IV+t" = inverse variance plus tau, which refers to random effects weighting according to the DerSimonian-Laird method.

In its current state (version 1.5) all results from the MIX program are identical (up to 4 decimals minimally) to results from the most recent versions of the metan, metabias, and metatrim commands in STATA. The small-study effect test by Macaskill[42] that was tested via STATA's regress command also turned out to be accurate. Table 5.4 and 5.5 are examples of the odds ratio validation results for data set 1^4.

Table 5.5. Summary of the validation with data set 1

	MIX	**STATA / CMA ***
Items in odds ratio meta-analysis	**Results**	**Results**
Fixed effect (IV) odds ratio		
OR (95% CI)	0.7677 (0.7196, 0.8190)	0.7677 (0.7196, 0.8190)
Z (*P* value)	8.0073 (< 0.0001)	8.01 (< 0.0001)
Q (*P* value)	39.48 (0.17)	39.48 (0.17)
I^2 (95% CI)	0.1895 (0 to 0.4749)	0.1895
Fail-safe N (tolerance level)	270 (175)	270 *
Rank correlation tau ** (*P* value)	0.1799 (0.141)	(0.141)
Egger's intercept (95% CI)	-0.2955 (-0.7880, 0.1970)	-0.2955 (-0.7880, 0.1970)
Macaskill's slope (*P* value)	0 (0.8396)	0 (0.840)
Trim and fill OR (95% CI)	0.7744 (0.7260, 0.8260)	0.774 (0.726, 0.826)
Trim and fill imputed studies	6	6
Random effects (IV + t^2) odds ratio		
OR (95% CI)	0.7619 (0.6825, 0.8506)	0.7619 (0.6825, 0.8506)
Z (P value)	-4.84 (< 0.0001)	-4.84 (< 0.0001)
t^2	0.117	0.117
Trim and fill OR (95% CI)	0.7828 (0.6907, 0.8871)	0.783 (0.691, 0.887)
Trim and fill imputed studies	6	6

This table shows the odds ratio validation of the inverse variance fixed effect and random effects analyses with data set 1. Though not shown here, the results from Mantel-Haenszel and Peto weighting were evaluated similarly and so were the other association measures (risk difference and risk ratio). The process was repeated for each data set. (*): Results marked with '*' were produced by CMA. (**): The rank correlation tau is a continuity corrected tau-b. Abbreviations: "IV" = inverse variance weighting, "OR" = odds ratio, "CI" = confidence interval, "IV+t" = inverse variance plus tau, which refers to random effects weighting according to the DerSimonian-Laird method.

With regard to the trim-and-fill analysis[43], the MIX program allows for calculations using the weighting method applied in the original meta-analysis, whereas both CMA and STATA use only fixed or random effects inverse variance methods when trimming and filling. While the calculations in MIX for trim-and-fill analyses with other weighting methods were verified manually and we have no reason to believe anything is wrong, we recommend using the inverse variance methods until more is known about approaches with alternative weighting.

Although we are in the process of completing a formal software comparison project, we are confident that the MIX program can compete in many respects (usability, analytical options, comprehensiveness, and export options) with most of the existing meta-analysis programs like Comprehensive Meta-Analysis[28], MetaWin[44], RevMan[45], or WEasyMA[46].

However, there are also still some limitations. One is the maximum number of studies that can be analyzed in the meta-analysis, which is now 100. Though systematic reviews finding 100 studies for analysis are still very rare, this is something that may change in the future. Furthermore, while sub-group analyses are easy to perform within MIX, they are currently not automated and during a sub-group analysis not all subgroups can be shown simultaneously in a single forest plot. The subgroup forest plot can however be created manually because the Excel graphs of individual forest plots are relatively easily formatted and stacked. We intend to improve the program with regard to these limitations in the near future.

Another important issue that we will focus on in upcoming updates is meta-regression. Although some univariable regression methods are integrated in the tests for small-study effects, the MIX program can currently not perform meta-regression. We realize that meta-regression, especially with multiple independent variables, is a valuable tool for assessing heterogeneity and adapting a meta-analysis accordingly, but it requires matrix calculations that are far more difficult to program in Excel or VBA than the standard tests. Currently, univariable meta-regression is possible with Comprehensive Meta-Analysis[28] and MetaWin[44]. However, like all dedicated meta-analysis packages they lack the option for multivariable meta-regression. We have started working on facilities for meta-regression within the MIX program and we hope it will be integrated sometime in 2007.

Finally, because we are still frequently updating the program and including new features, we have postponed the making of a hard-copy manual or methods guide until this process has stabilized.

5.4 Conclusion

The MIX program provides researchers, students, and lecturers with a free tool to perform state-of-the-art meta-analyses and learn or teach about what it is they are doing. It uses an innovative approach with Excel as a computing platform and even provides some numerical and graphical output that is not provided by other software. Results from version 1.5 of the MIX program are identical to those from

STATA, and MIX can be regarded as a comprehensive and valid tool for performing causal meta-analyses.

5.5 Availability and requirements

Project name:	MIX
Project homepage:	http://www.mix-for-meta-analysis.info
Operating system(s):	Microsoft Windows
Programming language:	Visual Basic (VB6, VBA)
Other requirements:	Microsoft Excel 2000 or later
License:	Open Source, free

Acknowledgments

The development and validation of the MIX program were supported by a one-year grant (#3042) from the Graduate School of Medical Sciences of Kitasato University. We are also grateful to all members of the Department of Medical Informatics of Kitasato University for the stimulating discussions during the project.

Conflict of interest

None declared.

References

1. Centre for Evidence Based Medicine - Oxford: Levels of Evidence and Grades of Recommendation. http://www.cebm.net/levels_of_evidence.asp.

2. Yusuf S, Cairns JA, Camm AJ, Fallen EL, Gersh BJ. *Evidence-Based Cardiology*. London: BMJ Publishing Group; 1998.

3. Yusuf S, Zucker D, Peduzzi P, et al. Effect of coronary artery bypass graft surgery on survival: overview of 10-year results from randomised trials by the Coronary Artery Bypass Graft Surgery Trialists Collaboration. *Lancet.* 1994;344(8922):563-570.

4. 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(4):248-254.

5. The Cochrane Collaboration. http://www.cochrane.org

6. Egger M, Davey Smith G, Altman D. *Systematic Reviews in Health Care: Meta-Analysis in Context*. London: BMJ Publishing Group; 2001.

7. Glasziou P, Irwig L, Bain C, Colditz G. *Systematic Reviews in Health Care: A Practical Guide*. Cambridge: Cambridge University Press; 2001.

8. Petitti DB. *Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine*. Second ed. Oxford: Oxford University Press; 2000.

9. Stangl D, Berry DA. *Meta-analysis in Medicine and Health Policy*. New York: Marcel Dekker; 2000.

10. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. *Methods for Meta-Analysis in Medical Research*. Chichester: Wiley; 2000.

11. Whitehead A. *Meta-Analysis of Controlled Clinical Trials*. Chichester: Wiley; 2002.

12. Egger M, Sterne JAC, Smith GD. Meta-analysis software. *BMJ*. 1998;316(7126):http://bmj.bmjjournals.com/archive/7126/7126ed7129.htm.

13. Normand SLT. Meta-analysis software - a comparative review - DSTAT, version 1.10. *American Statistician.* 1995;49:298-309.

14. Sterne JAC, Egger M, Sutton AJ. Meta-analysis software. In: Egger M, Davey Smith G, Altman DG, eds. *Systematic Reviews in Health Care: Meta-Analysis in Context*. 2nd ed. London: BMJ Books; 2001.

15. Sutton A, Lambert P, Hellmich M, Abrams K, Jones D. Meta-analysis in practice: A critical review of available software. In: Berry D, Stangl D, eds. *Meta-Analysis in Medicine and Health Policy*. New York: Marcel Dekker; 2000.

16. Bax L, Ikeda N. Explaining and performing common meta-analytic procedures in Japan: development of bilingual interactive software. Paper presented at: The 12th Cochrane Colloquium, 2004; Ottawa, Canada.

17. Bax L, Ikeda N, Shirataka M, Takeuchi A. Explaining common meta-analytic statistics in Japan with a simple Excel add-in. Paper presented at: The 24th Joint Conference on Medical Informatics, 2004; Nagoya, Japan.

18. Bax L, Tsuruta H, Ikeda N, Takeuchi A, Shirataka M. The MIX program, free software for learning, teaching, and exploring meta-analysis with Excel. Paper presented at: The 13th Cochrane Colloquium, 2005; Melbourne, Australia.

19. Bax L, Tsuruta H, Shirataka M, Takeuchi A, Ikeda N. The MIX program, an active way of learning about meta-analysis with Excel. Paper presented at: International Symposium: Systematic Review and Meta-Analysis, 2005; Wako, Japan.

20. Bax L, Yu L, Ikeda N, Tsuruta N, Moons K. *MIX: Comprehensive Free Software for Meta-analysis of Causal Research Data - Version 1.7. http://www.mix-for-meta-analysis.info*; 2008.

21. Knusel L. On the accuracy of statistical distributions in Microsoft Excel 2003. *Comput Stat Data Anal.* 2005;48(3):445-449.

22. McCullough BD, Wilson B. On the accuracy of statistical procedures in Microsoft Excel 2000 and Excel XP. *Comput Stat Data Anal.* 2002;40:713-721.

23. McCullough BD, Wilson B. On the accuracy of statistical procedures in Microsoft Excel 2003. *Comput Stat Data Anal.* 2005;49(4):1244-1252.

24. StataCorp. *Stata Statistical Software, Release 9.* College Station, TX: StataCorp LP; 2005.

25. Bradburn MJ, Deeks JJ, Altman DG. Metan - an alternative meta-analysis command (Metan 1.81). *Stata Tech Bull.* 2003;STB 44(sbe24):4-15.

26. Steichen TJ. Tests for publication bias in meta-analysis (Metabias 1.2.4). *Stata Tech Bull.* 2003;SJ3-4(sbe19_5):11.

27. Steichen TJ. Nonparametric trim and fill analysis of publication bias in meta-analysis (Metatrim 1.0.5). *Stata Tech Bull.* 2003;STB61(sbe39.2):11.

28. Borenstein M, Hedges L, Higgins J, Rothstein H. *Comprehensive Meta-Analysis Version 2*. Engelwood, NJ: Biostat; 2005.

29. Crowley P. Interventions for preventing or improving the outcome of delivery at or beyond term. *Cochrane Database Syst Rev.* 2000(2):CD000170.

30. Hodnett ED. Caregiver support for women during childbirth. *Cochrane Database Syst Rev.* 2000(2):CD000199.

31. Law MR, Wald NJ, Thompson SG. By how much and how quickly does reduction in serum cholesterol concentration lower risk of ischaemic heart disease? *BMJ*. 1994;308(6925):367-372.

32. Lightowler JV, Wedzicha JA, Elliott MW, Ram FS. Non-invasive positive pressure ventilation to treat respiratory failure resulting from exacerbations of chronic obstructive pulmonary disease: Cochrane systematic review and meta-analysis. *BMJ*. 2003;326(7382):185.

33. Pagliaro L, D'Amico G, Sorensen TI, et al. Prevention of first bleeding in cirrhosis. A meta-analysis of randomized trials of nonsurgical treatment. *Annals of Internal Medicine.* 1992;117(1):59-70.

34. Teo KK, Yusuf S, Collins R, Held PH, Peto R. Effects of intravenous magnesium in suspected acute myocardial infarction: overview of randomised trials. *BMJ*. 1991;303(6816):1499-1503.

35. Wahlbeck K, Cheine M, Essali MA. Clozapine versus typical neuroleptic medication for schizophrenia. *Cochrane Database Syst Rev.* 2000(2):CD000059.

36. Hodnett ED. Caregiver support for women during childbirth. *Cochrane Database Syst Rev.* 2000(2):CD000199.

37. Teo KK, Yusuf S, Collins R, Held PH, Peto R. Effects of intravenous magnesium in suspected acute myocardial infarction: overview of randomised trials. *BMJ*. Dec 14 1991;303(6816):1499-1503.

38. Crowley P. Interventions for preventing or improving the outcome of delivery at or beyond term. *Cochrane Database Syst Rev.* 2000(2):CD000170.

39. Lightowler JV, Wedzicha JA, Elliott MW, Ram FS. Non-invasive positive pressure ventilation to treat respiratory failure resulting from exacerbations of chronic obstructive pulmonary disease: Cochrane systematic review and meta-analysis. *BMJ.* Jan 25 2003;326(7382):185.

40. Wahlbeck K, Cheine M, Essali MA. Clozapine versus typical neuroleptic medication for schizophrenia. *Cochrane Database Syst Rev.* 2000(2):CD000059.

41. Law MR, Wald NJ, Thompson SG. By how much and how quickly does reduction in serum cholesterol concentration lower risk of ischaemic heart disease? *BMJ.* Feb 5 1994;308(6925):367-372.

42. Macaskill P, Walter SD, Irwig L. A comparison of methods to detect publication bias in meta-analysis. *Stat Med.* 2001;20(4):641-654.

43. Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics.* 2000;56(2):455-463.

44. Rosenberg MS, Adams DC, Gurevitch J. MetaWin: Statistical Software for Meta-Analysis Version 2. Sunderland, Massachusetts: Sinauer Associates; 2000.

45. The Nordic Cochrane Centre. Review Manager (RevMan). Version 4.2 for Windows. Copenhagen: The Cochrane Collaboration; 2003.

46. Chevarier P, Cucherat M, Freiburger T, et al. WeasyMA. Lyon: ClinInfo; 2000.

6

A SYSTEMATIC COMPARISON OF SOFTWARE FOR CAUSAL META-ANALYSIS

Based on:

Bax L, Yu LM, Ikeda N, Moons KG.

A systematic comparison of software dedicated to meta-analysis of causal studies.

BMC Med Res Methodol 2007, 7(40).

Abstract

Introduction. Computer software has become indispensable in meta-analysis and in the last decennia many programs have been developed. Our objective was to systematically assess the differences in features, results, and usability of currently available meta-analysis programs.

Methods. Systematic review of software. We did an extensive search on the Internet (Google, Yahoo, Altavista, and MSN) for specialized meta-analysis software. We included six programs in our review: Comprehensive Meta-Analysis (CMA), MetAnalysis, MetaWin, MIX, RevMan, and WEasyMA. Two investigators compared the features of the software and their results. Thirty independent researchers evaluated the programs on their usability while analyzing one data set.

Results. The programs differed substantially in features, ease-of-use, and price. Although most results from the programs were identical, we did find some minor numerical inconsistencies. CMA and MIX scored highest on usability and these programs also have the most complete set of analytical features.

Conclusion. In consideration of differences in numerical results, we believe the user community would benefit from openly available and systematically updated information about the procedures and results of each program's validation. The most suitable program for a meta-analysis will depend on the user's needs and preferences and this report provides an overview that should be helpful in making an informed choice.

6.1 Introduction

Computer software has become indispensable in meta-analysis and in the last decennia many programs have been developed. To aid potential users in choosing the software that fits their needs, there are a number of reviews and comparisons available[1-5]. The most recent one, however, dates back to five years ago and in the meantime the spectrum of available software has changed substantially. Also, most of the existing reviews have focused primarily on numerical features, such as which analytical models were available or what graphs could be produced. We believe that information on the validity or comparability of results and ease-of-use are equally important factors in the total applicability of the software. The purpose of the study described in this chapter was to systematically compare features, results, and usability of the currently available meta-analysis software, in particular those for causal meta-analysis.

6.2 Methods

6.2.1 Software search and selection

We decided, a priori, to focus on software that was solely dedicated to meta-analysis of randomized therapeutic or observational causal studies. General statistics packages were excluded. Furthermore, the software had to be actively maintained and supported, which was judged by either the time of the last software update (less than five years), bug report (less than five years), or website update (less than three years). We also decided to select only software with a graphical interface and mouse-click compatibility, which essentially excluded the DOS programs. Searches for software and publications related to their development and usage were done by two authors with combinations of the following keywords in Internet search engines of Google, Yahoo, AltaVista, and MSN: “meta-analysis”, “meta-analyses”, “systematic review”, “software”, “program”, “package”, “macro”, “add-in”, and “add-on”. The first search was done mid 2005 and the last search in June 2006. The software was purchased or downloaded if it appeared to fulfill the inclusion criteria provided above.

6.2.2 Assessment of numerical and graphical features

The assessment of the numerical and graphical features in the included meta-analysis programs was handled independently by two investigators and reviewed by all authors until there was consensus on all items. The programs were installed and tested on Windows XP and Windows 2000 systems in English and Japanese. Details of the documented features are provided in the tables of the results section.

6.2.3 Validity and comparability of meta-analysis

We searched the Internet and literature databases of medical and social sciences (PubMed, EmBase, Eric, and PsychInfo) for articles that reported validations of meta-analysis software. We also checked the website of each included program and made inquiries with its authors about their validation procedures.

Table 6.1 Data sets used in the validation

#	Author(s)	Date	Studies	Input type
1	Teo et al.[6]	1991	16	Group size, events
2	Wahlbeck et al.[7]	2000	11	Group size, mean, standard deviation
3	Pagliaro et al.[8]	1992	19	Association measure, standard error

The three data sets included did not involve extreme data with regard to heterogeneity, small-study effects, or rare events. The first data set did, however, have a study with no events in one study-arm and required a continuity correction for most analyses.

In addition to the search for validation reports, two reviewers actively investigated the comparability of the numerical results with data sets from three previously published[6,9,10] meta-analyses (Table 6.1). These data sets have been used as examples in methodological meta-analysis publications[5,11] and are representative of data sets commonly encountered in therapeutic or etiologic meta-analyses.

The first data set[6] contains per-group data from 16 randomized controlled trial articles with a dichotomous outcome, i.e. group sizes and event rates. One of the 16 included studies has no events in one of the treatment arms and the data set itself is subject to substantial small-study effects. The second data set[9] contains per-group data typically found in meta-analyses of controlled trials with a continuous outcome (group sizes, means, standard deviations). It contains data from 11 studies with heterogeneous results. The third data set[10] contains data as they could be found in meta-analyses of observational studies. The data are from 19 studies with a dichotomous outcome, like in the first data set. However, this time there are no per-group data available for each study but only the comparative association measures (odds ratios) and their standard errors.

For each data set, we compared the combined association measures, tests for heterogeneity, and tests for small-study effects (publication bias) derived from each of the studied meta-analysis programs. We focused on the most common association measures such as the risk difference, risk ratio, odds ratio, mean difference, Hedges' *g*, and Cohen's *d*, including their 95% confidence intervals. We used the metan (version 1.81)[12], metabias (version 1.4.2)[13], and metatrim (version 1.5.1)[13] programs of the general statistics software STATA[14] as reference in the software comparisons.

6.2.4 Assessment of usability

Finally, we performed a usability assessment amongst 30 researchers from various institutes and countries: Kitasato University (Japan), Tokai University (Japan), Utrecht University (the Netherlands), University of Amsterdam (the Netherlands), the Dutch Cochrane Center (the Netherlands), the University of Leuven (Belgium), and the Centre for Statistics in Medicine (UK). There were no specific inclusion criteria and the sample consisted of individuals from various departments and with various levels of experience with meta-analysis.

During the assessment sessions, participants were asked to install (evaluation versions of) each of the studied meta-analysis programs and to analyze one small data set of a meta-analysis with a dichotomous outcome (a shortened version of the previously described meta-analysis by Teo et al.[6]). As they completed this task, they scored the usability of each program in an electronic scoring list. This list (see appendix) was developed via a consensus session with (meta-analysis) experts from the disciplines of epidemiology, biostatistics, and medical informatics, who were asked which elements they considered important in meta-analysis software and what items they would use to judge its usability. The order in which each program was installed and assessed was determined by a computer generated randomization list and different for each participant.

6.3 Results

6.3.1 Included software

We found ten meta-analysis packages that were available for download or purchase via the Internet (Table 6.2). Half were no longer updated or had remained in their DOS stage and were excluded from our study. We included six programs in our comparison: Comprehensive Meta-Analysis (CMA) Version 2[15], WEasyMA 2.5[16], MetaWin 2.1[17], MetAnalysis[18], RevMan 4.2.8[19], and MIX 1.5[20]. Using less stringent inclusion/exclusion criteria did not change this software selection. Using more stringent criteria would exclude WEasyMA as various signs indicate that it may no longer be developed and supported.

Initially, our search did not pick up the still relatively unknown program called MetAnalysis. This software comes with a book and cannot be purchased separately. Neither the software nor the book is supported by a website, which is why we did not find it initially. At the time of inclusion, we could no longer assess it in the usability part, but have included it post-hoc in the assessment of comparability and features.

Table 6.2 Retrieved meta-analysis software

Software	OS requirements	Meta-analysis interface	Availability
Meta-Analysis 5.3	DOS	DOS menu	Free
EasyMA	DOS	DOS menu	Free
EpiMeta	DOS	DOS menu	Free
Meta-Stat	DOS	DOS menu	Free
Meta-Analyst	DOS	DOS menu	Free
CMA	WINDOWS	Graphical	Commercial
MetaWin	WINDOWS	Graphical	Commercial
WEasyMA	WINDOWS	Graphical	Commercial
RevMan	WINDOWS	Graphical	Free
MIX	WINDOWS	Graphical	Free

In the 'Meta-analysis interface' column, 'DOS menu' refers to a menu that has to be operated with the arrow keys (←↑→↓) and that can (mostly) not be manipulated with the mouse.

6.3.2 Numerical and graphical features

Below is a short summary of the numerical and graphical features in each of the reviewed programs; details are available from Tables 6.3 and 6.4. The following abbreviations are used: AM=association measure, CI=confidence interval, P=*P* value, W=weight, RD=risk difference, RR=risk ratio, OR=odds ratio, md=mean difference, HG=Hedges' *g*, CD=Cohen's *d*, CC=correlation coefficient, Z=Fisher's *Z*, IV=inverse variance weighting, MH=Mantel-Haenszel weighting, PETO=Peto's weighting, DL=Dersimonian & Laird weighting, Q=Cochran's or Breslow & Day's *Q*, I^2=Higgins's inconsistency statistic, t^2=between-study variance indicator, FSN=fail-safe number test, RC=rand correlation test, Egg=Egger's regression test, Mac=Macaskill's regression test, TF=trim-and-fill method, se=standard error, var=variance, N=sample size, TFP=trim-and-fill plot, HIST=histogram, NQ=normal quantile plot, BOX=box and whiskers plot.

Comprehensive Meta-Analysis (commercial software), also called CMA, has the highest profile in the Internet search engines of all included programs. It distinguishes itself from other programs by the option to enter effect sizes of different formats and the comprehensiveness of the numerical options and output. Data can be entered manually or via copy-and-paste in the CMA spreadsheet; direct import of text or other data files is not possible. The program features all major graphical presentations. The tutorial and manual are to-the-point and extensive. The program is actively maintained and the website is modern and regularly updated.

WEasyMA 2.5 (commercial software) stands out by the speed with which results become available after data set creation. Data cannot be imported or pasted and

need to be entered manually, cell by cell. Another limitation of this program is that it can only handle data from clinical trials with dichotomous outcomes, e.g. two-by-two table data. Although limited to these types of data, the program produces a wide variety of numerical and graphical output. The original author has indicated that the software is currently unsupported by a development team and may no longer be available in the near future.

MetaWin 2.1 (commercial software) is the only product that comes with a CD-ROM and a book. Other distinctive features are the effect size calculator, some uncommon graphs such as the normal quantile plot and a weighted histogram, and the option to use bootstrap confidence intervals. The interface resembles a spreadsheet program and various data files can be imported. For some changes in the analysis, data range selections have to be repeated, which is somewhat more time-consuming compared to methods used by other programs. In contrast to most other software, all calculations are based on *t*-distributions and boot-strap methods are also available. The help files and the book are extensive and detailed.

MetAnalysis 1.0 (commercial software) is not sold separately, but comes as a bonus feature of a book[18]. Similar to WEasyMA, it is limited to studies with descriptive data on dichotomous outcomes. Data cannot be pasted or imported and must be entered manually, cell by cell. Once the data are entered and the calculations performed, numerical data can be produced in a print preview screen and graphs in separate windows. A nice feature is the radial part of the Galbraith plot, which is lacking in most other software. The software also has the facilities to enter loss to follow-up / drop-out information and use the studies in the meta-analyses with per-protocol or intention-to-treat analysis. The software does not contain help files and does not have a website, but the book makes up for that to some extent.

RevMan 4.2.8 (free for private and academic use) was developed by the Cochrane Collaboration. It stands out due to its extensive features for collaborative management of systematic reviews. The analytical functions of the program cannot be accessed without first creating a review structure and because import and copy-and-paste functionality are also limited, getting started requires more preparation than with most other software. Once data are in the analysis module, analysis is straightforward. Output is detailed, though without tests for publication bias and no other graphs than the forest and funnel plot. The help resources in RevMan are extremely thorough.

MIX 1.5 (free software) is a recently developed program. Its most prominent features are the comprehensive graphical output, detailed numerical options, built-in data sets corresponding to those in a number of books, and extensive tutor functions. MIX is the only program that will not function by itself and it requires Microsoft Excel 2000 or later to run. Another limitation is the maximum number of data sets, which is currently 100. Data sets can be created by manual input as well as by importing text files or Excel workbooks. The numerical and graphical options are diverse and comprehensive.

6.3.3 Meta-analysis results

Our Internet and database search did not yield any publications on the validity or validation of any of the programs, except for MIX[20]. Authors of all programs were contacted to determine whether (yet unpublished) evidence of validation procedures was available. Authors of RevMan indicated that validation data were made public via notes and abstracts at Cochrane Collaboration meetings and conferences. The authors of CMA, MetAnalysis, and MetaWin stated that all procedures had been checked extensively with external programs, spreadsheets, and occasionally by hand, though had not been made public. For CMA, Excel sheets with such data are available upon request. We received no information on validation procedures from the authors of WEasyMA.

We found no discrepancies in meta-analysis results between STATA, MIX, and RevMan. In CMA, we found a small inconsistency in results of publication bias tests, but this was corrected via an update while we were writing this article.

MetaWin's results were different from STATA's results (and thus also from results in CMA, MIX, and RevMan) because MetaWin mostly uses a *t*-distribution where the aforementioned programs use a *z*-distribution. We did, however, find what seemed to be a terminological inconsistency, as the Mantel-Haenszel labeled method used in MetaWin for odds ratio analyses gave results that were identical to those from Peto's method in the other programs (albeit with confidence limits based on a *t*-distribution).

Since MetAnalysis and WEasyMA can only analyze data from two-by-two tables, the comparability assessments were limited to one data set[6]. Analyses in MetAnalysis were very similar though not always identical to those from STATA. We found that if we entered experimental group data first (as is the case in all other software), an incorrect event coding is applied that causes the software to calculate risk differences and odds ratios of survival even if mortality is entered as event. For risk differences this only changes the sign, but for odds and odds ratios

Table 6.3 Meta-analysis software - basic feature comparison

	CMA	WEasyMA	MetaWin	MetAnalysis	RevMan	MIX
General						
URL	*meta-analysis.com*	*weasyma.com*	*metawinsoft.com*	-	*cc-ims.net/RevMan*	*mix-for-meta-analysis.info*
Corporate user price	$1295.00	$490.00	$150.00	$73.00	$650	Free
Student user price	$395.00	$280.00	$ 75.00	$ 73.00	Free	Free
Download / program size	30 Mb	3 Mb	9 Mb	5 Mb	9 Mb	20 Mb / 50 Mb
Compatibility	Windows	Windows	Windows	Windows	Windows	Windows
Last update	2006	2002	2002	2005	2005	2006
License	Single user	Single user	Single user	Single user	Open (non-commercial	Open
Input options						
Manual input	✓	✓	✓	✓	✓	✓
Copy & paste	✓		✓		(✓)	✓
Text file import			✓			✓
File import (Excel, other software)			✓			✓
Descriptive dichotomous, e.g. n(total), n(y=1)	✓	✓	✓	✓	✓	✓
Descriptive continuous, e.g. n, m, sd	✓		✓		✓	✓
Comparative, e.g. theta, se/var	✓		✓		✓	✓
Multi-format (mixed in one data set)	✓					
Single data input / selection	✓	✓	(✓)	✓	✓	✓
Maximum number of studies	Unlimited	Unlimited	Unlimited	Unlimited	Unlimited	100
Information sources						
Within-program HTML help		(✓)	✓		✓	(✓)
Printable manual	✓		✓	✓	✓	
Description of methods/calculations			✓	(✓)	✓	(✓)
Additional information sources (PDFs/tutorials)	✓		✓			✓
Up-to-date website	✓	×	✓		✓	✓
Export options						
Copy output to clipboard	✓	✓	✓	✓	✓	✓
Export to office application(s)	✓					
Report creation	✓			✓		✓
Setting copy file type (e.g. bmp, jpg or wmf)		✓	✓			✓

The '✓' indicates the presence and no mark indicates the absence of a feature. The '(✓)' means that the feature is limited or partially in development, and the '×' means it was not working correctly at the time of our assessments.

Table 6.4 Meta-analysis software - analytical feature comparison

	CMA	WEasyMA	MetaWin	MetAnalysis	RevMan	MIX
Computational setting options						
Number of decimals	✓	✓	✓			✓
Alpha level / confidence intervals	✓	✓	✓		✓	✓
'Add 0.5' method continuity correction	✓	✓	✓	✓	✓	✓
'Inverse group size' method continuity correction						✓
Variance for mean differences	✓					✓
Variances for standardized mean differences	✓					✓
Bootstrap confidence intervals			✓			
Numerical output						
Individual study data	AM, CI, P, W, other	AM, CI, W, other	AM, other	AM, CI, other	AM, CI, P, W, other	AM, CI, P, W, other
Association measures – risk	RD, RR, OR	RD, RR, OR	RD, RR, OR	RD, OR	RD, RR, OR	RD, RR, OR
Association measures – means & standardized measures	MD, HG, CD, other		HG, other		MD, HG	MD, HG, CD
Association measures – other	CC, Z					
Fixed effect models / weighting	IV, MH, PETO	IV, MH, PETO,	IV, MH, PETO	IV, MH, PETO	IV, MH, PETO	IV, MH, PETO
Random effects models / weighting	DL	DL	DL	DL	DL	DL
Cumulative analyses	Several variables	Several variables	Several variables	(✓) Only graph		Several variables
Heterogeneity	Q, I^2, t^2	Q	Q	Q, I^2	Q, I^2	Q, I^2, t^2,other
Small study effect / publication bias	FSN, RC, EGG, TF	EGG	FSN, RC	FSN, EGG		FSN, RC, EGG, MAC,
Meta-regression	Single moderator	×	Single moderator			
Graphical output						
Forest plot	✓	✓	✓	✓	✓	✓
- Points proportional to weights	✓				✓	✓
- Annotations in rows possible	✓			✓	✓	✓
- Cumulative possible	✓			✓		✓
Funnel plot (1/se, se, var, N, P)	1/se, se	1/se, se, N	var, N	N	1/se	1/se, se, N, P
Galbraith plot		✓	✓	✓ (radial)		✓
Exclusion sensitivity plot	✓					✓
Trim and fill plot	✓					✓
L'Abbe plot		✓		✓		✓
Other plots			HIST, NQ			BOX, HIST, NQ, other
Graph formatting	✓	✓	✓	✓	✓	✓

The '✓' indicates the presence and no mark indicates the absence of a feature. The '(✓)' means that the feature is limited or partially in development, and the '×' means it was not working correctly at the time of our assessments.

it gives the reciprocal of the intended results[21]. Although the book mentions that control data are to be entered in the first data column, this is contrary to conventions in all other software that we reviewed and we feel it is not unlikely that users make this mistake. The software has currently no built-in guard against this.

In WEasyMA, we found that if there is a study with zero events in one study arm, the program uses a continuity correction that adds 0.5 to all two-by-two table cells of all studies. This initially caused large differences with the results from STATA's metan (in which by default 0.5 is added only to the cells of the study with the zero event study arm). We attempted to reproduce WEasyMA's results by using the same continuity correction in STATA, but some association measures and confidence intervals of individual studies were still different. The WEasyMA authors were contacted but did not respond to our request for clarification of the discrepancies.

6.3.4 Usability

Of the 30 participating researchers, 26 provided quantitative data that were suitable for analysis (Table 6.5). Trouble with the electronic user form or installation of software made the data from four researchers incomplete and they were excluded from the quantitative part.

Table 6.5 Meta-analysis software usability ratings

Items	MIX	CMA	MetaWin	RevMan	WEasyMA
All researchers (26)					
Overall (min, max)	8.6 (6.7, 10)	6.9 (3.7, 9.7)	6.2 (4.3, 8.7)	6.1 (4.3, 8.3)	4.2 (1, 7.3)
Getting started	8.6	7.4	6.8	7.6	4.5
Data preparation	8.3	6.3	6.3	4.5	2.6
Usability in analysis	8.8	7.1	5.6	6.3	5.9
Experienced (7)					
Overall (min,max)	8.1 (7.0, 9.7)	6.8 (6.0, 7.3)	5.9 (4.3, 7.7)	5.4 (4.3, 6.3)	3.3 (1, 5.7)
Getting started	8.0	7.6	6.2	7.5	2.8
Data preparation	8.3	6.3	6.3	3.0	2.0
Usability in analysis	8.0	6.6	5.4	6.3	5.3
Inexperienced (19)					
Overall (min, max)	8.7 (6.7, 10)	7 (3.7, 9.7)	6.3 (4.3, 8.7)	6.3 (4.7, 8.3)	4.6 (1.3, 7.3)
Getting started	8.8	7.3	6.9	7.7	5.0
Data preparation	8.3	6.3	6.3	5.0	2.8
Usability in analysis	9.1	7.3	5.6	6.3	6.1

MIX scored highest on the overall usability (8.6), followed by CMA (6.9), MetaWin (6.2), RevMan (6.1), and WEasyMA (4.2). Users were, in general, least satisfied with the procedures for data preparation. Stratifying the results by user

experience, country, or familiarity with a certain program did not reveal any specific trends. Installation of WEasyMA and CMA was troublesome for some researchers. Qualitative statements mostly concerned problems with the installation (WEasyMA, CMA), error messages in French (WEasyMA), and difficulties with data set creation (WEasyMA, RevMan). Favorable comments included praise for the user interfaces (MIX, RevMan, CMA), help system (RevMan), speed of analysis (WEasyMA), and within-program tutoring (MIX, CMA).

6.4 Discussion

Meta-analysis is an indispensable tool in current-day synthesis of research data from multiple studies, and systematic reviews with meta-analyses occupy the top position in the hierarchy of evidence. Software for meta-analysis has evolved over the years and available reviews are relatively outdated. We therefore considered it timely to provide a systematic overview of the features, criterion validity, and usability of the currently available software that is dedicated to meta-analysis of causal (therapeutic and etiologic) studies. Compared to existing reviews[1-5], it contains more detailed information on the merits and demerits of the available programs and follows a more systematic approach.

We studied four commercial programs (CMA, WEasyMA, MetaWin, and MetAnalysis) and two free programs (RevMan and MIX). The features of the commercial programs were not necessarily more extensive than those of the free ones. In particular MIX stood out in terms of numerical options and graphical output. CMA was generally most versatile, in particular in options for analysis of various types of data. With regard to the comparability of results, MIX, RevMan, and CMA produced numerical results that were identical to results from STATA's metan, metabias, and metatrim. However, the CMA program required an update to get these results. MetaWin's results are different and slightly more conservative, with confidence intervals based on a *t*-distribution or bootstraps. WEasyMA may produce results that are not valid and we discourage the use of this program for scientific purposes until a validated update is available. MetAnalysis should be used with care as data have to be entered manually and in the correct order. With regard to the latter, users do not receive a warning and the subsequent results can be invalid. The usability survey shows that preparing data for analysis is the hardest part in each program. MIX and CMA were seen as most user-friendly and WEasyMA scored less favorably in this respect. The latter program also appears to be no longer supported by a development team and its website has gone off-line during the writing of this manuscript.

Our comparison has been limited to software dedicated to meta-analysis only and does not include general statistics packages. The primary reason to leave them out was because they are structurally very different, making direct comparisons inappropriate. Central to this issue is software syntax: most general packages require thorough knowledge of their syntax in order to produce and alter graphs that are common in meta-analysis; the dedicated packages, however, produce such graphs with a few or sometimes even a single click.

In addition, the syntax knowledge required to do more advanced meta-analyses with the general packages means that in a usability survey all participants would have to be expert statisticians, capable of writing and adapting syntax for meta-analysis in all major general software packages. This is not only not feasible in the current setting, it would also make the participating individuals no longer representative of the (sometimes relatively inexperienced) users of the software in the scientific and academic community.

Due to the lack of a gold standard, we resorted to between-program comparisons and a criterion validation with STATA's user-written commands metan, metabias and metatrim as reference. Our choice for STATA was based on its versatility and explicit use in two major books on meta-analysis[5,11]. We realize that STATA itself is also user-written and potentially subject to similar validity issues than the other programs. The fact that CMA, MIX, and RevMan produced results that were identical to results from STATA, at least with the three data sets we selected, justifies to some extent our use of STATA as a reference standard.

The results of our usability survey should be regarded as a rough indication. First, the number of participants was still relatively small. Furthermore, it is not unlikely that there may be some bias in favor of RevMan because some users were already familiar with this program. Sensitivity analysis (leaving these participants out), however, did not reveal such a trend. MetAnalysis could unfortunately not be assessed as it was included after the start of the usability assessment. A point regarding MIX is that it was created following a development focus list[20] that was created in a similar fashion to our usability scoring list. Assessment of both lists reveals that a number of items are very similar. Though this indicates that the lists are indeed reflecting the demands of statistical software users, it also means that the MIX program was likely to do well in our assessment. One could argue, however, that any program that is systematically developed to satisfy its users' demands should perhaps deservedly score high.

An important point to which we would like to draw attention is the lack of accessible public information about the manner in which meta-analysis programs have been validated. Only the website of the MIX program includes specific information about this and MIX is the only program with a peer-reviewed and published validation report[20]. Without such reports, authors, reviewers, editors, and consumers of evidence have no reference for judgments about the suitability of the software for scientific purposes. This is of course equally applicable to the user-written meta-analysis macros for general statistics software. We advocate more rigor and transparency in this area.

Finally, we are aware that the world of information technology changes constantly and by the time this manuscript is published, it is possible that some updates have become available or that new products have been launched. We apologize beforehand for our lack of timing. Like a traditional review, we intend to update this investigation in due time. In conclusion, the most suitable meta-analysis software for a user depends on his or her demands; no single program may be best for everybody. The information provided in this article, in particular the data in Table 6.3 and 6.4, should give users the opportunity to make a substantiated decision.

Acknowledgments

This study was not supported by any particular grant. The authors would like to express their gratitude to all researchers who participated in the usability assessment sessions.

Conflict of interest

None of the authors have financial conflicts of interest. The first author is, however, the primary developer of one of the free programs (MIX) studied in this review. The other authors were co-authors of an introductory article about MIX. To reduce personal biases, all tasks were handled by multiple investigators and the subjective usability assessments were assigned (by study design) to individuals other than the authors.

References

1. Arthur W, Bennett W, Huffcutt A. Choice of software and programs in meta-analysis research: Does it make a difference? *Educ Psychol Meas.* 1994;54:776-787.

2. Egger M, Sterne J, Smith G. Meta-analysis software. *BMJ*. 1998;316(7126): Website only: *http://bmj.bmjjournals.com/archive/7126/7126ed7129.htm*.

3. Normand S. Meta-analysis software - a comparative review -DSTAT, version 1.10. *Amer Statistician.* 1995;49:298-309.

4. Sterne J, Egger M, Sutton A. Meta-analysis software. In: Egger M, Davey Smith G, Altman D, eds. *Systematic Reviews in Health Care: Meta-Analysis in Context*. 2nd ed. London: BMJ Books; 2001.

5. Sutton A, Lambert P, Hellmich M, Abrams K, Jones D. Meta-analysis in practice: A critical review of available software. In: Berry D, Stangl D, eds. *Meta-Analysis in Medicine and Health Policy*. New York: Marcel Dekker; 2000.

6. Teo KK, Yusuf S, Collins R, Held PH, Peto R. Effects of intravenous magnesium in suspected acute myocardial infarction: overview of randomised trials. *BMJ*. Dec 14 1991;303(6816):1499-1503.

7. Wahlbeck K, Cheine M, Essali MA. Clozapine versus typical neuroleptic medication for schizophrenia. *Cochrane Database Syst Rev.* 2000(2):CD000059.

8. Pagliaro L, D'Amico G, Sorensen TI, et al. Prevention of first bleeding in cirrhosis. A meta-analysis of randomized trials of nonsurgical treatment. *Ann Intern Med.* Jul 1 1992;117(1):59-70.

9. Wahlbeck K, Cheine M, Essali MA. Clozapine versus typical neuroleptic medication for schizophrenia. *Cochrane Database Syst Rev.* 2000(2):CD000059.

10. Pagliaro L, D'Amico G, Sorensen TI, et al. Prevention of first bleeding in cirrhosis. A meta-analysis of randomized trials of nonsurgical treatment. *Ann Intern Med.* Jul 1 1992;117(1):59-70.

11. Egger M, Davey Smith G, Altman D. *Systematic reviews in health care: meta-analysis in context.* London: BMJ Publishing Group; 2001.

12. Bradburn M, Deeks J, Altman D. Metan - an alternative meta-analysis command (Metan 1.81). *Stata Tech Bull.* 2003;STB 44(sbe24):4-15.

13. Steichen T. Tests for publication bias in meta-analysis (Metabias 1.2.4). *Stata Journal.* 2003;SJ3-4(sbe19_5):11.

14. StataCorp. *Stata statistical software, Release 9*. College Station, TX: StataCorp LP; 2005.

15. Borenstein M, Hedges L, Higgins J, Rothstein H. *Comprehensive Meta-Analysis Version 2*. Engelwood, NJ: Biostat; 2005.

16. Chevarier P, Cucherat M, Freiburger T, et al. *WeasyMA*. Lyon: ClinInfo; 2000.

17. Rosenberg M, Adams D, Gurevitch J. *MetaWin: Statistical Software for Meta-Analysis Version 2*. Sunderland, Massachusetts: Sinauer Associates; 2000.

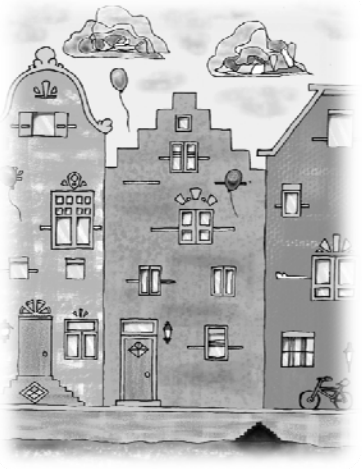
18. Leandro G. *Meta-analysis in Medical research*: Blackwell Publishing, BMJ Books; 2005.

19. The Nordic Cochrane Centre. *Review Manager (RevMan). Version 4.2 for Windows*. Copenhagen: The Cochrane Collaboration; 2003.

20. Bax L, Yu L, Ikeda N, Tsuruta H, Moons K. Conference proceeding: Validation of a freely available and comprehensive meta-analysis add-in for excel. *0393-2990.* 2006;21(supplement):58.

21. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Stat Med.* Jun 15 2002;21(11):1575-1600.

APPLICATIONS



7

UNCERTAIN EFFECTS OF ROSIGLITAZONE

Based on:

Diamond G, Bax L, Kaul S.

Uncertain effects of rosiglitazone on the risk for myocardial infarction and cardiovascular death.

Ann Intern Med 2007, 147(8):578-582.

Abstract

Introduction. A recent, widely publicized meta-analysis of 42 clinical trials concluded that rosiglitazone was associated with an approximately 43% increased risk for myocardial infarction and an approximately 64% increased risk for cardiovascular death. The sensitivity of these conclusions to several methodological choices was not assessed. The meta-analysis was not based on a comprehensive search for all studies that might yield evidence about rosiglitazone's cardiovascular effects. Studies were combined on the basis of a lack of statistical heterogeneity, despite substantial variability in study design and outcome assessment. The meta-analytic approach that was used required the exclusion of studies with zero events in the treatment and control groups.

Methods. A meta-analysis of randomized clinical trials. Existing meta-analysis data were re-analyzed with alternative methods.

Results. Alternative meta-analytic approaches that use continuity corrections show lower odds ratios that are not statistically significant.

Conclusion. We conclude that the risk for myocardial infarction and death from cardiovascular disease for diabetic patients taking rosiglitazone is uncertain: neither increased nor decreased risk is established.

7.1 Introduction

7.1.1 The rosiglitazone analysis

In June 2007, the New England Journal of Medicine (NEJM) published a meta-analysis of 42 clinical trials involving 27,847 patients that ignited a firestorm of controversy by charging that treatment with rosiglitazone (Avandia®, GlaxoSmithKline, Brentford, United Kingdom), a widely prescribed peroxisome proliferator activated receptor γ agonist, was associated with about a 43% greater risk of myocardial infarction and a 64% greater risk of cardiovascular death than placebo or other antidiabetic regimens[1].

In performing their analysis, the investigators screened 116 phase 2, 3, and 4 trials. Of these, 48 met the predefined inclusion criteria of having a randomized comparator group and at least 24 weeks of drug exposure in all groups. Six of the 48 trials, with an unknown number of patients, were excluded because they did not report any cardiovascular events. Of the remaining 42 trials (38 double-blind and four open-label), only ten assessed rosiglitazone monotherapy against placebo. Of the 32 trials comparing rosiglitazone with other antidiabetic therapy, 28 evaluated rosiglitazone versus placebo as an add-on therapy to sulfonylurea (n=12), metformin (n=10), insulin (n=5), or usual care (n=1), and four compared rosiglitazone monotherapy head-to-head with sulfonylurea or metformin. Thus, most of the trials were placebo-controlled.

7.1.2 Limitations of the analysis

The investigators acknowledged several limitations to their analysis. Misclassification and ascertainment errors were possible, because most studies were designed to assess end points other than cardiovascular disease. Study protocols had variable inclusion and exclusion criteria and drug dosing regimens. Overall event rates were low, partially because trial durations were relatively short, ranging from 24 to 52 weeks. Patient-level data were unavailable for time-to-event analysis. Relevant events, such as stroke or noncardiovascular death, were not reported, and whether reported death and myocardial infarction events were mutually exclusive was not always clear. In addition, the investigators did not present details of their literature search method apart from accessing published literature, trial registries, and U.S. Food and Drug Administration summary data. Without this systematic review context, the comprehensiveness of the analyzed data set is difficult to judge.

These matters aside, we scrutinize two additional limitations. First, study designs and populations were heterogeneous, yet data were pooled based on lack of

statistical heterogeneity as assessed by Cochran's *Q* test. This test has limited ability to detect variation across studies with sparse data[2]. We believe that the decision to pool all studies despite design and population heterogeneity likely led to artificial inflation and precision of the risk estimate. Concordantly, the investigators' own subgroup analyses that were limited to the small trials alone or to the two large trials (DREAM and ADOPT) did not demonstrate statistically significant associations.

Furthermore, one might reasonably question whether results from the three trials that targeted patients with Alzheimer's disease (*n*=1) or psoriasis (*n*=2) who did not have diabetes should be combined with results from other trials that included patients with diabetes or prediabetes. Because rosiglitazone is already contraindicated in patients with heart failure, one might also reasonably limit the assessment of risk to patients without that contraindication and not combine data from the single study in patients with diabetes who had congestive heart failure with data from other studies. Incidentally, this trial exhibited the highest number of myocardial infarctions (*n*=5) and cardiovascular deaths (*n*=3) among all the small trials in the rosiglitazone treatment group.

Second, a single methodological approach, Peto's fixed-effects model, was used to combine data. The authors justified this choice by the absence of statistically significant heterogeneity and the overall paucity of cardiovascular events, and also referenced a simulation study by Bradburn et al.[3]. This simulation exercise found the Peto method performed well when events were sparse and numbers of patients in the study arms were balanced. In the meta-analysis at hand, several studies had major imbalances whereby numbers of patients assigned rosiglitazone were two to three times greater than the numbers of patients assigned the comparator. In such cases, the Peto method is reported to perform less satisfactorily[3,4], and alternative methods, such as the Mantel-Haenszel method with a study arm continuity correction and other approaches, with and without continuity corrections, merit consideration[4-7].

The methodological choices about what and how to combine are particularly salient when no events are observed in both the treatment and comparator groups (zero total event trials). The authors excluded 4 such trials from the myocardial infarction analysis and 19 such trials from the cardiovascular death analysis. It is common practice to exclude all zero total event trials from meta-analyses because they provide no information about the magnitude of the odds or risk ratios and do not contribute to producing a combined treatment effect greater or less than nil[3,4,8].

On the other hand, we and others think that zero total event trials may provide relevant information by showing that event rates for both the intervention and control groups are low and relatively equal[5-7]. Including such trials can sometimes decrease the effect size estimate and narrow confidence intervals. Imagine a collective experience of 200,000 patients, equally divided into treatment and control arms, in which no events have been observed. Our best estimate of the event rate based on this experience is clearly less than 1 in 100,000. Imagine then that a single event is found in a subsequent trial with 100 patients. An analysis that ignored all the prior exculpatory information would estimate the risk to be 1%—at least 1000 times greater than it may actually be. Moreover, if we were to do a meta-analysis that ignored zero total event trials, the small trial would not be weighted in relation to all evidence available, and would receive a larger weight than we would naturally assume.

We think that similar concerns about ignoring relevant data apply to trials exhibiting zero events in only one or the other study group (zero-event-trials). Consider the equation used to calculate the pooled estimate of the odds ratio according to Peto:

$$OR_{Peto} = \exp\left(\sum_{j=1}^{k} (O_j - E_j) / \sum_{j=1}^{k} V_j \right),$$

where O_j is the observed event rate in the treatment group, E_j is the expected event rate in the treatment group, and V_j is the variance of the difference $(O_j - E_j)$ for the j^{th} of k trials. According to this equation, the Peto odds ratio is inflated if the number of trials with zero events in the control group is greater than that of trials with zero events in the treatment group. In the data analyzed by Nissen and Wolski, this situation occurred with a ratio of 20:6 for myocardial infarction and 15:2 for cardiovascular death.

7.2 Methods

We re-analyzed the data set of 42 trials considered by Nissen and Wolski using a meta-analytic software package called MIX[9-11]. We estimated the pooled odds ratio as measure of association using fixed effects (e.g. Mantel-Haenszel) and random-effects (DerSimonian-Laird) models. When applicable, we used methods with or without two continuity corrections. One is a constant correction (CC) that adds values of 0.5 to all cells of the two-by-two contingency table of the study that was selected for correction. The other is a study arm correction (SAC) that adds values proportional to the reciprocal of the size of the opposite study arm. The corrections were applied in two scenarios: (1) to studies with zero event in one

arm, excluding studies with zero events in both arms (CC and SAC), and (2) to all studies with either zero event in one or both arms (CC+ or SAC+), effectively including all studies. Studies without zero events received no correction.

Table 7.1 describes the continuity corrections. The CC for continuity adds 0.5 to each cell of the contingency table, increasing the treatment and control arm sizes by 1 and the total study sample sizes by 2. In study C in Table 7.1, the event rate in the treatment arm becomes 1.5 and the non-event rate 399.5. The event rate and non-event rate in the placebo arm are increased accordingly, to 0.5 and 100.5. The SAC for continuity adds a value proportional to the reciprocal of the size of the opposite study arm, normalized to a sum of 1 for event and no-event cells, also resulting in an increase in the total study sample size by 2. With R being the ratio of group sizes and S being the sum of corrections (set to 1 here), the SAC adds a factor of R/S*(R+1) to the larger arm and 1/S*(R+1) to the other arm[4]. Applying this to study C in Table 7.1, R is 400/100=4 and the correction in the larger arm is 4/1*(4+1)=4/5=0.8 and in the smaller arm 1/1*(4+1)=1/5=0.2. The latter correction has been shown to be less biased in the case of study arm imbalances[4].

Table 7.1 Examples of continuity corrections

Correction *	Treated +	Treated -	Placebo +	Placebo -	Odds ratio calculation	Odds ratio
Study A	1	249	0	250	(1/249)/(0/250)	NA
CC	1+0.5	249+0.5	0+0.5	250+0.5	(1.5/249.5)/(0.5/250.5)	3.01
CC+	1+0.5	249+0.5	0+0.5	250+0.5	(1.5/249.5)/(0.5/250.5)	3.01
SAC	1+0.5	249+0.5	0+0.5	250+0.5	(1.5/249.5)/(0.5/250.5)	3.01
SAC+	1+0.5	249+0.5	0+0.5	250+0.5	(1.5/249.5)/(0.5/250.5)	3.01
Study B	0	250	0	250	(0/250)/(0/250)	NA
CC	-	-	-	-	-	Excl.
CC+	0+0.5	250+0.5	0+0.5	250+0.5	(0.5/250.5)/(0.5/250.5)	1
SAC	-	-	-	-	-	Excl.
SAC+	0+0.5	250+0.5	0+0.5	250+0.5	(0.5/250.5)/(0.5/250.5)	1
Study C	1	399	0	100	(1/400)/(0/100)	NA
CC	1+0.5	399+0.5	0+0.5	100+0.5	(1.5/399.5)/(0.5/100.5)	0.75
CC+	1+0.5	399+0.5	0+0.5	100+0.5	(1.5/399.5)/(0.5/100.5)	0.75
SAC	1+0.8	399+0.8	0+0.2	100+0.2	(1.8/399.8)/(0.2/100.2)	2.26
SAC+	1+0.8	399+0.8	0+0.2	100+0.2	(1.8/399.8)/(0.2/100.2)	2.26
Study D	0	400	0	100	(0/400)/(0/100)	NA
CC	-	-	-	-	-	Excl.
CC+	0+0.5	400+0.5	0+0.5	100+0.5	(0.5/400.5)/(0.5/100.5)	0.25
SAC	-	-	-	-	-	Excl.
SAC+	0+0.8	400+0.8	0+0.2	100+0.2	(0.8/400.8)/(0.2/100.2)	1

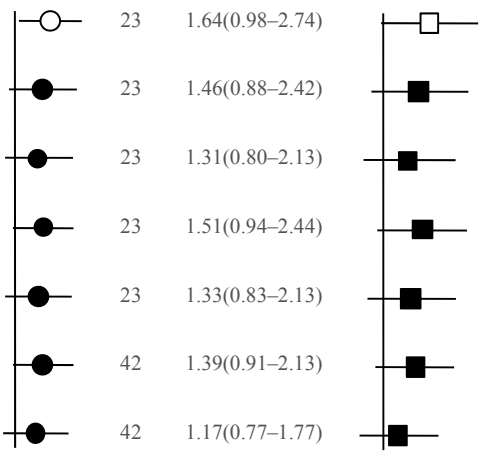
(*) The four corrections are: (1) constant correction for trials with zero events in one study arm (CC), (2) constant correction for studies with zero events in one or more study arms (CC+), (3) study arm dependent correction for trials with zero events in one study arm (SAC), (4) study arm dependent correction for trials with zero events in one or more study arms (SAC+).

7.3 Results

Table 7.2 shows the pooled odds ratios and 95% CIs for analysis of all 42 trials. The odds ratios ranged from 1.43 to 1.26 for myocardial infarction and from 1.64 to 1.17 for cardiovascular death. Models without continuity correction yielded larger odds ratios than continuity-corrected models, with the CCs providing slightly lower estimates than the SACs. Although the odds ratios from these analyses were elevated, the CIs all contained an odds ratio of unity. Moreover, they suggest greater uncertainty than was reported in the original report but do not rule out the possibility that rosiglitazone increases risk for myocardial infarction or cardiovascular death. The caterpillar plot next to Table 7.2 illustrates this with the odds ratios (squares for myocardial infarction and circles for cardiovascular death) and their confidence intervals plotted in relation to each other and to an odds ratio of unity (the vertical line).

Table 7.2 Meta-analytic odds ratios for myocardial infarction and cardiovascular death

	Myocardal infarction		Cardiovascular death	
Method	*k*	OR (95% CI)	*k*	OR (95% CI)
Fixed,Peto	38	1.43(1.03–1.98)	23	1.64(0.98–2.74)
Fixed,IV(SAC+)	38	1.34(0.97–1.84)	23	1.46(0.88–2.42)
Fixed,IV(CC+)	38	1.29(0.94–1.76)	23	1.31(0.80–2.13)
Fixed,MH(SAC)	38	1.36(1.00–1.84)	23	1.51(0.94–2.44)
Fixed,MH(CC)	38	1.28(0.95–1.72)	23	1.33(0.83–2.13)
Fixed,MH(SAC)	42	1.35(1.00–1.82)	42	1.39(0.91–2.13)
Fixed,MH(CC)	42	1.26(0.93–1.69)	42	1.17(0.77–1.77)



k – number of studies; OR – odds ratio; IV – inverse variance; MH – Mantel-Haenszel.

Subgroup analyses exhibited similar odds ratios but wider confidence intervals than our primary analysis, consistent with the smaller sample sizes (see the online Table at www.annals.org). None of these analyses conclusively adjudicate the association between rosiglitazone and the risk for myocardial infarction or cardiovascular death in particular groups of patients.

The controversy engendered by Nissen and Wolski's analysis caused the investigators of the large, ongoing RECORD (Rosiglitazone Evaluated for Cardiac

Outcomes and Regulation of Glycaemia in Diabetes) trial to perform an unplanned interim analysis[12]. This trial is designed as a noninferiority comparison of add-on rosiglitazone to metformin or sulfonylurea. The interim analysis showed a hazard ratio of 0.83 (95% CI, 0.50 to 1.36) for the adjudicated secondary end point of cardiovascular death and 1.17 (CI, 0.75 to 1.82) for the adjudicated secondary end point of myocardial infarction among the 4447 trial patients. On the basis of these interim results, the investigators concluded that there was no evidence of any increase in mortality rate but that data were insufficient to determine whether there was an increase in the risk for myocardial infarction. The statistical power of these comparisons was limited because of an unexpectedly low event rate and incomplete follow-up (a mean of 3.7 years instead of the planned median of 6 years), but the confidence intervals overlapped those reported in Nissen and Wolski's meta-analysis. This observation, along with a more than 30% excess risk for myocardial ischemic events in rosiglitazone-treated patients reported in the meta-analysis conducted by GlaxoSmithKline, lends continued support to those questioning the cardiovascular safety of rosiglitazone[13,14]. A recent Cochrane review of 18 studies in patients with diabetes indicated a tendency toward increased risk for myocardial infarction with rosiglitazone treatment but could not confirm statistically significant differences in odds ratios for rosiglitazone versus controls[15]. A large observational study in 33,363 patients did not show an increased risk for adverse cardiovascular outcomes in patients taking rosiglitazone compared with other therapies[16]. Overall, these reports indicate confusing and sometimes conflicting results about cardiovascular risk associated with rosiglitazone therapy.

7.4 Discussion

The risk for myocardial infarction and death from cardiovascular disease for diabetic patients taking rosiglitazone is uncertain. Neither increased nor decreased risk is established. Using the same data as was analyzed in a recent, widely publicized meta-analysis[1], we showed the fragility of effect sizes for the above risks. We think that excluding trials with zero events in the index meta-analysis probably exaggerated risk estimates and that including these trials by applying continuity adjustments in this instance temper the exaggerated estimates.

Our analysis is restricted by the same limitations as those in the index analysis: short follow-up durations, low event rates, absence of patient-level data about time-to-event, variable and probably incomplete outcome ascertainment, and inability to reliably assess total mortality or composite outcomes such as death or myocardial infarction. Neither analysis is a comprehensive systematic summary of all available evidence regarding the potential cardiovascular risks of rosiglitazone.

We acknowledge that our analysis does not establish the amount of bias associated with different analytic methods for pooling trials with sparse events or with various choices of continuity corrections. We did not test other choices for continuity corrections. Lower correction values than those commonly used and applied in our analysis might cause less of a shift whereas higher correction values would result in a greater shift towards a null effect (odds ratio of unity) compared to unadjusted pooled estimates.

In the end, we believe that only prospective research designed for the specific purpose of establishing the cardiovascular risk of rosiglitazone will resolve the controversy about its safety. In our opinion, available evidence does not justify what the authors of the original meta-analysis (as well as the media, the U.S. Congress, and worried patient groups) decried as an "urgent need for comprehensive evaluations".

Conflict of interest

None declared.

References

1. Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *N Engl J Med.* Jun 14 2007;356(24):2457-2471.
2. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ.* Sep 6 2003;327(7414):557-560.
3. Bradburn MJ, Deeks JJ, Berlin JA, Russell Localio A. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Stat Med.* Jan 15 2007;26(1):53-77.
4. 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.* May 15 2004;23(9):1351-1375.
5. Cook DJ, Witt LG, Cook RJ, Guyatt GH. Stress ulcer prophylaxis in the critically ill: a meta-analysis. *Am J Med.* Nov 1991;91(5):519-527.
6. Sankey SS, Weissfeld LA, Fine MJ, Kapoor W. An assessment of the use of the continuity correction for sparse data in meta-analysis. *Commun Stat Simul Comp.* 1996;25:1031-1056.
7. Friedrich JO, Adhikari NK, Beyene J. Inclusion of zero total event trials in meta-analyses maintains analytic consistency and incorporates all available data. *BMC Med Res Methodol.* 2007;7:5.

8. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. *Stat Med.* Nov 1991;10(11):1665-1677.

9. Bax L, Yu LM, Ikeda N, Moons KG. A systematic comparison of software dedicated to meta-analysis of causal studies. *BMC Med Res Methodol.* 2007;7:40.

10. Bax L, Yu LM, Ikeda N, Tsuruta H, Moons KG. Development and validation of MIX: comprehensive free software for meta-analysis of causal research data. *BMC Med Res Methodol.* 2006;6:50.

11. Bax L, Yu LM, Ikeda N, Tsuruta N, Moons KG. *MIX: Comprehensive Free Software for Meta-analysis of Causal Research Data - Version 1.7.* http://www.mix-for-meta-analysis.info; 2008.

12. Home PD, Pocock SJ, Beck-Nielsen H, et al. Rosiglitazone evaluated for cardiovascular outcomes - an interim analysis. *N Engl J Med.* Jul 5 2007;357(1):28-38.

13. Psaty BM, Furberg CD. The record on rosiglitazone and the risk of myocardial infarction. *N Engl J Med.* Jul 5 2007;357(1):67-69.

14. Nathan DM. Rosiglitazone and cardiotoxicity - weighing the evidence. *N Engl J Med.* Jul 5 2007;357(1):64-66.

15. Richter B, Bandeira-Echtler E, Bergerhoff K, Clar C, Ebrahim SH. Rosiglitazone for type 2 diabetes mellitus. *Cochrane Database Syst Rev.* 2007(3):CD006063.

16. McAfee AT, Koro C, Landon J, Ziyadeh N, Walker AM. Coronary heart disease outcomes in patients receiving antidiabetic agents. *Pharmacoepidemiol Drug Saf.* Jul 2007;16(7):711-725.

8

NEUROMUSCULAR ELECTRICAL STIMULATION AND MUSCLE STRENGTH

Based on:

Bax L, Staes F, Verhagen A.
Does neuromuscular electrical stimulation strengthen the quadriceps femoris? A systematic review of randomized controlled trials.
Sports Med. 2005;35(3):191-212.

Abstract

Introduction. Devices for neuromuscular electrical stimulation (NMES) are increasingly used by individuals without specific injuries and are standard equipment in most physical therapy practices. The most often stimulated muscle group is the quadriceps femoris. We designed a systematic review and meta-analysis of randomized controlled trials to determine whether NMES is an effective modality for strength augmentation of the quadriceps femoris.

Methods. A full content search for randomized controlled trials was performed in Medline, Embase, Cinahl, the Cochrane Controlled Trials Register, and the Physical Therapy Evidence Database. Maximum volitional isometric or isokinetic muscle torque in Nm was used as main outcome measure.

Results. Thirty-five trials were included and evaluated. A fundamental distinction was made between the trials using subjects with unimpaired quadriceps femoris muscles and the trials using post-injury or post-operative subjects. In the unimpaired quadriceps subgroup, meta-analyses were performed for the comparisons 'NMES versus no exercises' and 'NMES versus volitional exercises'. All other comparisons were evaluated descriptively. The included trials were generally of poor quality and meta-analytic data indicate that publication bias may be present. The evaluated data suggest that, both for the unimpaired and impaired quadriceps, NMES makes sense compared with doing no exercises but volitional exercises appear to be more effective in most situations.

Conclusion. Based on the available evidence, NMES may only be preferred over volitional training for within-cast muscle training and perhaps in specific situations where volitional training does not receive sufficient patient compliance. Further research should be directed toward identifying the clinical impact at activity and participation levels and the optimal stimulation parameters of this modality.

8.1 Introduction

Applying an electrical current to neuromuscular tissue to trigger muscle contractions has been in practice since the 18th century[1]. When used on weakened but otherwise healthy muscles, the application of the electrical current to the neuromuscular junction and surrounding muscle fibers causes the muscle to contract and this type of muscle training is often referred to as neuromuscular electrical stimulation (NMES).

The general objective of NMES is to improve fundamental, training-related muscle properties (intra-muscular blood-flow, maximum force output, force endurance) through repetitive contractions, and from this viewpoint NMES is not only applied in various rehabilitative settings, but also in athletic or preventive strength training[2,3]. Unfortunately, the abundance of muscle stimulators in clinics appears to contrast sharply with the availability and quality of the available evidence for their efficacy.

Randomized controlled trials (RCTs) generally provide the most valid methodology to evaluate the efficacy of therapeutic interventions, and systematic reviews of RCTs are assumed to provide the highest level of evidence in therapeutic clinical research[4]. Although some narrative reviews related to NMES are available[5-9], systematic reviews are rare[10], and have not focused on randomized controlled trials.

As we expected substantial heterogeneity and many trials, we considered several methods to narrow down our subject. Eventually, focusing only on the muscle group that is most often stimulated in clinical practice and most often investigated in clinical trials appealed to all of us, as it would not limit sensitivity or subgroup analyses based on stimulation parameters and patient characteristics. This muscle group turned out to be the quadriceps femoris[5,7-9]. Anticipating residual heterogeneity, we a priori defined a small number of subgroups, the most important of which being studies of subjects with unimpaired quadriceps muscles and studies of (orthopedic) post-operative or post-injury subjects.

Since electrical stimulators capable of eliciting muscle contractions in one form or another are increasingly being sold directly to patients, clinicians should be aware of the evidence supporting their use. This article attempts to provide a systematic and, where possible, quantitative overview of current research data from RCTs. The main clinical question was: “Could neuromuscular electrical stimulation in general be an effective modality for increasing the strength of the quadriceps femoris in adults?”

8.2 Methods

8.2.1 Search and selection of eligible studies

The in- and exclusion criteria are summarized in Table 8.1. The language of the trial report, publication date, and journal type were not part of the in- or exclusion criteria. Searches were conducted in Medline, Embase, Cinahl, Cochrane Controlled Trials Register, and the Physical Therapy Evidence Database. The last search was performed in July 2002.

Table 8.1 The inclusion and exclusion criteria

Inclusion criteria
- Trials that involve a comparison of at least two interventions, one of which should be a form of electrical stimulation with the objective to cause physiological changes in the muscle by means of the elicited contractions. The other arm of the trial should be a no-treatment control, a sham treatment or a volitional exercise treatment.
- Trials that involve random allocation or intended-to-be-random allocation of eligible people to intervention groups to receive or not to receive one or more interventions that are being compared. Intended-to-be-random allocation means methods of allocation such as alternation or allocation by odd and even birth dates or hospital record numbers.
- Trials that use participants who are older than 18 (no maximum age limit).
- Trials that use maximum voluntary isometric strength or maximum voluntary isokinetic strength as primary or secondary outcome measure.
Exclusion criteria
- Cross-over trials in which each subject is subjected to more than one intervention.
- Trials in which the sample is drawn from a population of individuals with physical or psychological disorders affecting senso-motoric functioning.

A search filter containing a methodological component, a population component, an intervention component, and an outcome component was developed for the searches. The methodological component was based on commonly used search terms for identifying RCTs[11]. The complete filter is presented in Table 8.2.

One reviewer performed the initial search and first selection. These procedures included the screening of the reference lists of the preliminarily selected reports and background articles. During the more detailed selection, two reviewers read the complete articles and reached full consensus on in- and exclusion decisions.

8.2.2 Quality assessment

A previously validated criteria-list was used to assess the overall general quality of the included RCTs[12]. The list was developed via extensive Delphi consensus procedures and will in this review be referred to as the Delphi list. It contains items of internal validity (randomization procedures, concealment of

randomization, baseline comparability, blinding of assessors/care providers/patients, intention-to-treat analysis), generalizability (description of eligibility criteria), and effect quantification (description of point estimates and variability). The total number of items is nine and in this review they contributed equally to the summary score, which is displayed as a percentage of the maximum score.

Table 8.2 The search filter with four components in PubMed-search format

Methodological component
"(randomized controlled trial[pt] OR controlled clinical trial[pt] OR randomized controlled trials OR random allocation OR double blind method OR single blind method OR clinical trial[pt] OR clinical trials OR (clin*[tw]AND trial*[tw]) OR ((singl*[tw] OR doubl*[tw] OR trebl*[tw] OR tripl*[tw]) AND (blind*[tw] OR mask*[tw])) OR placebos OR placebo*[tw] OR random*[tw] OR research design OR comparative study OR evaluation studies OR follow up studies OR prospective studies OR control OR controlled OR prospective*[tw] OR volunteer*[tw])"
Population component
"NOT ((animal[mh]) NOT (human[mh] AND animal[mh])) NOT (cerebrovascular accident[mh] OR cerebrovascular disorders[mh] OR hemiplegia[mh] OR quadriplegia[mh] OR paraplegia[mh] OR paralysis[mh] OR incontinence)"
Determinant component
"AND ((electric stimulation therapy[mh] OR electric*[tw]) AND (muscle OR muscular) AND (stimulat*[tw] OR train*[tw] OR exerci*[tw] OR contracti*[tw]))"
Outcome component
"AND ((muscle OR muscular) AND (strength*[tw] OR forc*[tw] OR contracti*[tw]) AND (isokinetic OR isometric OR eccentric))

Three reviewers performed independent quality assessments, with disagreement to be resolved by consensus. Three external reviewers, in the process of testing a similar quality assessment list, provided independent advice when necessary.

8.2.3 Data extraction

Two independent reviewers performed the data extraction with regard to study domain, interventions, and outcomes using standardized forms. The main outcome was isometric torque, measured by maximum volitional contractions (MVC) in Newton-meter. The focus was on the (end of training) outcome measures of which the test procedures were similar to the training procedures. If data were not presented numerically but in graphs, these graphs were copied and enlarged to A3 size to extract the data. Outcome measures on an activity or participation level[13] were documented as well. At this stage, the studies were allocated to the a-priori subgroups, most importantly: studies of subjects with unimpaired quadriceps muscles versus studies of post-operative or post-injury subjects, and studies using independent NMES ($NMES^{-}$) versus superimposed NMES ($NMES^{+}$). In the impaired quadriceps subgroup, a further distinction was made between post-immobilization stimulation and during-immobilization stimulation.

8.2.4 Data analysis

In the unimpaired quadriceps subgroup, meta-analyses could be performed for the comparisons 'NMES versus no exercises' and 'NMES versus volitional exercises' with mean differences weighted by the inverse of the variance. All other comparisons were evaluated descriptively. Whenever references are made to significance levels or confidence intervals an alpha-level of 0.05 was used, corresponding with 95% confidence intervals. Since substantial (clinical) heterogeneity was anticipated and the data quality and quantity did not suffice for meta-regression, a random effects model was used and sensitivity analyses were performed with regard to the alternative use of a fixed effect model. When possible, mean changes from baseline were used instead of the mean of the post-training follow-up, which allowed for incorporation of baseline differences in the effect estimation. Unfortunately, not all trials provided sufficient data for mean change from baseline calculations, and of the studies that did only three provided variance estimates of the mean changes. We therefore decided to use the variances of the follow-up scores for all included studies to ensure comparable weighting in the meta-analysis.

Clinical heterogeneity was evaluated descriptively by exploring the differences between the RCTs with regard to study population, types of interventions, and outcomes. When applicable, statistical heterogeneity was tested formally by a Chi-square test and informally by eyeballing the overlap of the 95% confidence intervals in the forest plots. We used a graphical method to explore sources of statistical heterogeneity[14]. In this method, a scatter plot is constructed with the contribution of a trial to the overall Cochran *Q* statistic for heterogeneity on the x-axis and the contribution to the overall effect on the y-axis. The trials that have a large influence on the overall effect and contribute substantially to the heterogeneity can be identified in the upper-right quadrant of the graph.

Within the meta-analyses, we performed funnel plot analyses with the means on the x-axis and the standard error on the y-axis to assess publication bias. To assess the potential impact of any 'missing trials', identified by the asymmetry of the funnel plot, we used the trim-and-fill method[15] with data imputation to make the funnel plot symmetrical and to assess the influence of the imputed trials on the overall result. Additional one-way sensitivity analyses were done for a-priori defined validity items and statistical methods.

8.3 Results - part 1: search and selection

After completing the search procedures, 2297 citations were reviewed and 69 trials were selected for further screening. Eventually, 35 studies were included.

Reasons for exclusion were: (a) no randomization (*n*=17), (b) outcome measures not quadriceps strength (*n*=11), and (c) no control group used (*n*=6). The full flow of trials is shown in Figure 8.1.

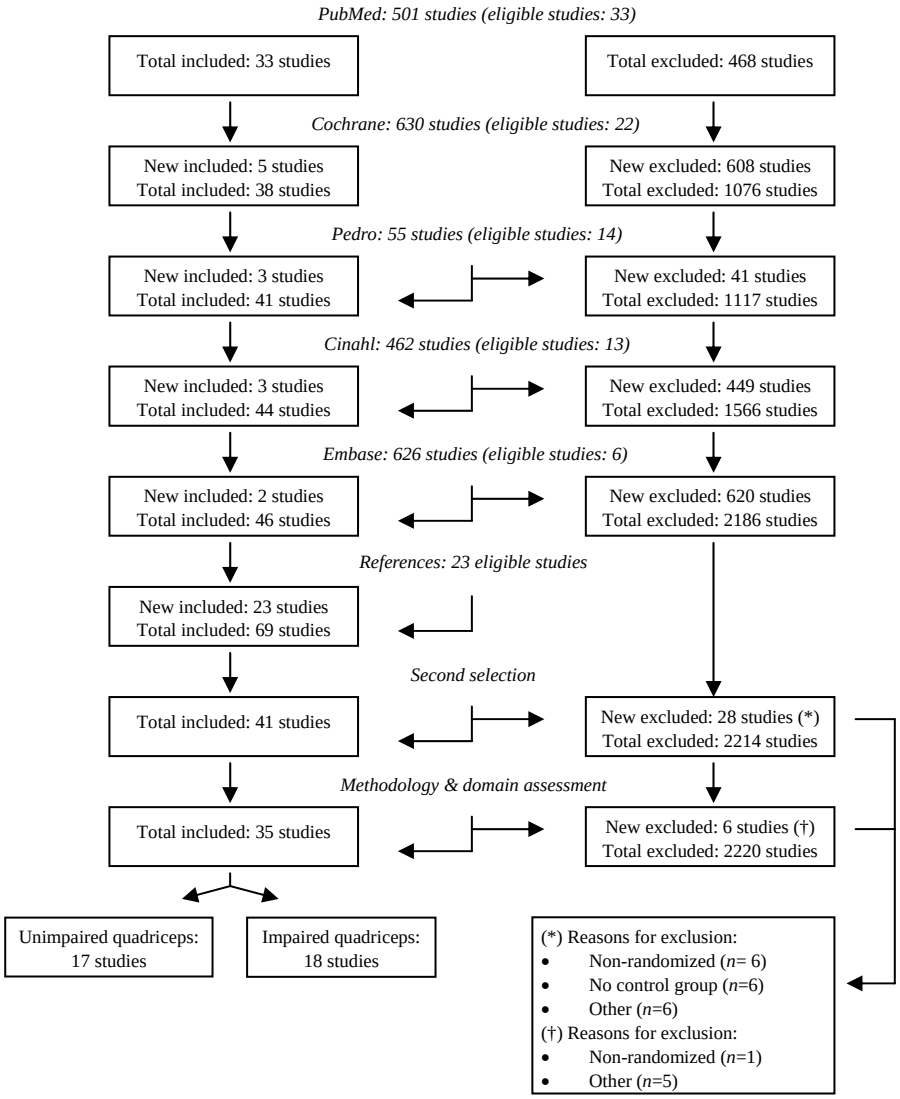


Figure 8.1 Flow of trials in the review

Table 8.3 provides detailed information about the search results per database. It gives the initial pool of studies that was retrieved by the search engine (A), the number of studies from this pool that were eventually included (B), and an evaluation of the search in terms of sensitivity and efficiency. The sensitivity and

efficiency of a search procedure is quantified in this table as follows: 'sensitivity' = included studies found in database divided by total number of included studies, and 'efficiency' = included studies found in database divided by number of retrieved and screened studies from database.

Table 8.3 Quantification of search results

Databases	Retrieved (A)	Included (B)	Sensitivity (B_i / B_{total})	Efficiency (B_i / A_i)
Total	-	35	-	-
(a) PUBMED	501	20	57%	3.9%
(b) COCHRANE	630	18	51%	2.9%
(c) PEDRO	55	10	29%	18.2%
(d) CINAHL	462	11	31%	2.4%
(e) EMBASE	626	4	11%	0.6%
(f) REFERENCES	23	6	17%	-
(a) + (b)	1131	24	69%	2.1%
(a) + (b) + (c)	1186	28	80%	2.4%
(a) + (b) + (f)	1154	30	86%	2.6%
(a) + (b) + (c) + (f)	1209	32	91%	2.6%

PubMed provided the most sensitive search environment: had only the PubMed database been used, 20 of the 35 finally included studies would have been included. A combination of PubMed, Pedro, Cochrane Library, and a reference list check would have retrieved 32 of the 35 articles (91%). The Pedro database provided only 10 of the 35 articles (29%), but this database provided the most efficient search environment (18% of the retrieved and screened trials were amongst the finally included trials). The high efficiency of the search in this database can be explained by the fact that this database is mainly oriented towards physical therapy.

8.4 Results - part 2: adults with unimpaired quadriceps femoris

Of the 17 included trials, five trials had only male subjects[3,16-19], five trials had only female subjects[20-24], and seven trials had both male and female subjects[25-31]. Of the trials that reported age means of the subjects, the overall mean age was 28.0 (SD: 8.3). The average number of subjects per study arm of the trials that could be included in a quantitative analysis was ten (range: 6-20). Only one study used outcome measures at the activity level in addition to muscle torque[3]. With regard to the intervention, most studies used isometric training methods with the knee flexed at 30 to 90 degrees. The average training period lasted four to five weeks. Various stimulation frequencies, pulse widths, and training regimes were applied. Table 8.4 describes the study details with regard to methods, study domain, interventions, and results.

8.4.1 Quality assessment

The Delphi scores ranged from 22% to 56%, with a mean of 43% (SD: 12%). The percentage of items of the Delphi list that were answered with 'don't know' was 29%. The trials scored extremely low on the allocation concealment item and the blinding items. The Delphi scores (end scores as well as per-item scores) for each study are included in the third row of Table 8.4, and the order of the answers ('Y' or 'N' or '?') corresponds with the order of the items of the Delphi list[12].

Table 8.4 Characteristics of the included studies with unimpaired subjects

ID and Subjects	Methods *	Bias †	NMES features
Balogun, 1993 Healthy men, mean age 22 (*n*=38)	3-arm RCT Other	Delphi score 44% Y-?-Y-Y-?-?-?-Y-N	Dynamic concentric, no carrier, twin peak pulse, (a) 80 Hz, (b) 45 Hz, (c) 20 Hz, 65-75 µs, hold 10 s, rest 50 s, 10 reps, 3 days/week, 6 weeks
Caggiano, 1994 Healthy aged men (*n*=22), mean age 72	2-arm RCT NMES– vs VOL	Delphi score 33% Y-?-?-Y-?-N-N-Y-N	Isometric, no carrier, 25-50 Hz, ~110 µs, ramp 5s, hold 5s, rest 50s, 10 reps, 3 days a week, 4 weeks
Currier, 1983 Healthy men and women (*n*=42), age 18-32.	4-arm RCT NMES$^{-/+}$ vs NO NMES$^{-/+}$ vs VOL NMES^{-} vs NMES^{+}	Delphi score 44% Y-?-Y-Y-?-N-N-Y-N	Isometric, superimposed, independent, 2500 Hz carrier, 50 Hz excitatory frequency, ~110 µs, ramp 5s, hold 10s, rest 50 s, 10 reps, 3 days a week, 5 weeks
Fahey, 1985 M: Healthy men F: Healthy women (*n*=28), m. age 27)	3-arm RCT NMES^{-} vs NO Other	Delphi score 44% Y-?-Y-Y-?-N-N-Y-N	Isometric, no carrier, 50 Hz, ramp 2 s, hold 8 s, rest 5 s, 15 minutes, 3 days per week, 6 weeks
Hortobagyi, 1998 Healthy women (*n*=24), age 22-26	3-arm RCT NMES^{-} vs NO NMES^{-} vs VOL	Delphi score 44% Y-?-Y-Y-?-N-N-Y-N	Eccentric, 2500 Hz carrier, 50 Hz excitatory frequency, 4x per week, 6 weeks
Hortobagyi, 1999 Healthy women (*n*=32), mean age 25	4-arm RCT NMES^{-} vs NO NMES^{-} vs VOL Other	Delphi score 22% Y-?-?-Y-?-N-N-N-N	Eccentric, 2500 Hz carrier, 50 Hz excitatory frequency, output 38-67 mA, 4x per week, 6 weeks
Kramer, 1983 Healthy women (*n*=40), mean age 21	4-arm RCT NMES$^{-/+}$ vs NO NMES$^{-/+}$ vs VOL NMES^{-} vs NMES^{+}	Delphi score 56% Y-?-Y-Y-?-N-N-Y-Y	Isometric, superimposed *and* independent no carrier, 100 Hz, 500 µs (?), hold 10 s, rest 50 s, output maximally tolerable, 10 reps, 1 set, 2-3x per week, 4-5 weeks
Kubiak, 1987 Healthy men and women (*n*=29), age 18-30.	3-arm RCT NMES^{-} vs NO NMES^{-} vs VOL	Delphi score 44% Y-?-?-Y-?-N-N-Y-Y	Isometric, 2500 Hz carrier, 50 Hz excitatory frequency, ramp up 5 s, hold 10s, rest 50 s, output 45%-134% MVIC, 10 reps, 1 set, 3x per week, 5 weeks
Lai, 1988 Healthy men and women (*n*=24), age 23-26	3-arm RCT NMES^{-} vs NO Other	Delphi score 56% Y-?-Y-Y-?-N-N-Y-Y	Isometric, (a) <25% MVIC, (b) >50% MVIC), no carrier, 50 Hz, 200 µs, hold 5 s, rest 5 s, 10 reps, 3 sets (1 min rest), 5x per week, 3 weeks

* 'NMES^{-}', independent NMES; 'NMES^{+}', superimposed NMES; 'NO', no exercise group; 'VOL', volitional exercise. † Scores on Delphi items: 'Y', yes; 'N', no; '?', don't know.

Table 8.4 Characteristics of the included studies with unimpaired subjects (cont.)

ID and Subjects	Methods *	Bias †	NMES features
Laughman, 1983 Healthy men and women (*n*=58), age 21-39.	3-arm RCT NMES^{-} vs NO NMES^{-} vs VOL	Delphi score 56% Y-?-Y-Y-?-N-N-Y-Y Analysis conform randomization	Isometric, 2500 Hz carrier, 50 Hz excitatory frequency, ramp-up 5 s, hold 10 s, rest 50 s, 10 reps, 1 set, 5x per week, 5 weeks
Maffiuletti, 2000 Healthy male basketball players (*n*=20), mean age 25	2-arm RCT NMES^{-} vs NO	Delphi score 56% Y-?-Y-Y-?-N-N-Y-Y Analysis conform randomization	NMES^{-} asunct, isometric, no carrier, 100 Hz, 400 µs, hold 3 s, rest 17 s, output 60-100 mA, 48 reps (30 minutes), 3x per week, 4 weeks
McMiken, 1983 Healthy men and women (*n*=16), age 19-27	2-arm RCT NMES^{-} vs VOL	Delphi score 33% Y-?-?-Y-?-N-N-Y-N	Isometric, no carrier, 75 Hz, 100 µs, hold 10 s, rest 50 s, output 70-80% MVIC, 10 reps, 1 set, 3x per week, 3 weeks
McMiken, 1983 Healthy men and women (*n*=16), age 19-27	2-arm RCT NMES^{-} vs VOL	Delphi score 33% Y-?-?-Y-?-N-N-Y-N	Isometric, no carrier, 75 Hz, 100 µs, hold 10 s, rest 50 s, output 70-80% MVIC, 10 reps, 1 set, 3x per week, 3 weeks
Mohr, 1985 Healthy women (*n*=18), age 21-29	3-arm RCT NMES^{-} vs NO NMES^{-} vs VOL	Delphi score 44% Y-?-Y-Y-?-N-N-Y-N	Isometric, no carrier, 50 Hz, 17-28 µs, ramp-up 3.3 s, hold 10 s, rest 10 s, output maximally tolerated, 10 reps, 1 set, 5x per week, 3 weeks
Romero, 1982 Healthy women (*n*=23), mean age 23	2-arm RCT NMES vs NO	Delphi score 44% Y-?-Y-Y-?-N-N-Y-? Drop-outs > 10%, no additional analysis	Isometric, 2000 Hz (carrier?), hold 4 s, rest 4 s, ~50 mA, 15 min, 2x per week, 5 weeks
Selkowitz, 1985 Healthy men and women (*n*=24), age 18-32	2-arm RCT NMES^{-} vs NO	Delphi score 56% Y-?-Y-Y-?-N-N-Y-Y Analysis conform randomization	Isometric, 2200 Hz carrier, 50 Hz excitatory frequency, 450 µs, hold 10 s, rest 120 s, output at maximally tolerated level (28-90 mA), 10 reps, 1 set, 3x per week, 4 weeks
Singer, 1983 Healthy men, mean age 31.8 (*n*=25)	5-arm RCT NMES^{-} vs NO NMES^{-} vs VOL Other	Delphi score 22% Y-?-N-Y-?-N-N-N-?	Hold 10s, rest 50 s, output up to 80 mA, 10 reps, 3x per week, 4 weeks
Wolf, 1986 Healthy males athletes, (*n*=27), age 24-52.	3-arm RCT NMES^{+} vs NO (NMES as adjunct)	Delphi score 22% Y-?-N-Y-?-N-N-N-?	Dynamic, 75 Hz excitatory frequency, output maximally tolerable, 24 sessions (6 weeks), only last 12 sessions (3 weeks) with EMS

* 'NMES^{-}', independent NMES; 'NMES^{+}', superimposed NMES; 'NO', no exercise group; 'VOL', volitional exercise. † Scores on Delphi items: 'Y', yes; 'N', no; '?', don't know.

None of the studies described procedures to conceal the randomization to the individuals responsible for the in- and exclusion of the participants (concealed allocation). Six trials had no dropouts or loss to follow-up[3,22,27-29,31]. All other trials used per-protocol analyses. In some trials with a relatively large number of dropouts or loss to follow-up, i.e. > 10%, and a relatively small sample size[16,24,25], this type of analysis may have introduced some bias. None of the trials used blinding methods of any kind, but as far as description of the eligibility is concerned, all trials scored positively. However, with regard to the intervention, only four trials[16,22,25,31] provided the NMES parameters in sufficient detail to allow

for reproduction of the treatment. In three cases[18,19,21] the description of results for the relevant outcomes were insufficient for inclusion in the meta-analysis.

8.4.2 NMES versus no exercises

The studies that did not report the results sufficiently enough for usage of the data in the meta-analyses[18,19,21] were all applicable to this comparison. The study by Hortobagyi et al. [21] reports significantly better results in the NMES$^-$ group than in the no-exercise group. The other two trials by Singer et al. [18] and Wolf et al. [19] report no significant differences. None of these trials used allocation concealment, blinding methods, or methods to analyze or correct for potential systematic error in the results due to dropouts or non-compliance. In the trial by Hortobagyi et al. [21], the training and measurement were based on eccentric contractions, whereas the other trials used isometric methods. Wolf et al. investigated the effect of NMES as an adjunct to basic exercise programs, which may also explain the fact that NMES did not result in significant strength improvement compared to the control group. The results of the remaining studies[3,20,22-29,31] were included in the meta-analyses, with one reference being included twice because the men and women were randomized separately[26].

The weighted mean difference for the 'independent NMES versus no exercises' comparison is 8.00 Nm (95% CI: 2.79, 13.21) and is based on 12 trial references and 235 subjects. Both the Chi-square value (χ^2=32.82; *df*=11; *P*=0.0006) and the eyeballing of the overlap of the 95% confidence intervals of the trials clearly indicate heterogeneity. Nevertheless, since all point estimates of the trials favor NMES$^-$, the heterogeneity is more relevant to the extent to which NMES$^-$ is more effective than to whether NMES$^-$ is more effective at all. One trial[3] with basketball players assessed two outcome measures at the activity level, namely vertical jump height from a squatted position and after a so-called counter-movement. The results were measured in cm and the between-group mean difference was 4.9 cm (95% CI: 3.4, 6.4) for the squat-jump and 0.3 cm (95% CI: -1.0, 1.6) for the counter-movement jump, both subtly favoring the NMES$^-$ group.

Only two studies were included in the 'superimposed NMES versus no exercises' subgroup [22,25]. Both studies indicate that NMES$^+$ may be more effective than no exercises. The 95% confidence interval of the point estimate in the study by Kramer and Semple[22] does not include the null value, but the confidence interval of the point estimate in the study by Currier and Mann[25] does. The weighted mean difference for this comparison is 25.61 Nm (95% CI: 9.47, 41.75) and is based on 38 subjects. The overall pooled point estimate based on both subgroups was 10.15 Nm (95% CI: 4.71, 15.58).

Both subgroups were included in a graphical assessment of the statistical heterogeneity (Figure 8.2). Two studies stand out[3,20], which may be related to the measurement of the results in N instead of Nm[20], the eccentric training protocol[20], or the highly trained and therefore probably highly motivated participants[3]. Deleting the trials from the meta-analysis does not substantially alter the pooled point estimate.

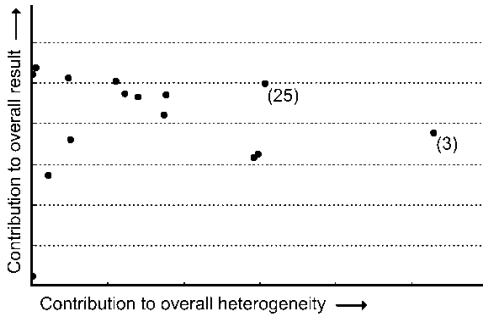


Figure 8.2 Heterogeneity of NMES versus no exercises – healthy quadriceps

The numbers in the figure correspond with the reference numbers used throughout the text.

The forest plot and funnel plot of both subgroups together are presented in Figure 8.3. The asymmetry of the funnel plot indicates that publication bias should not be ruled out. Imputation of six trials with the trim-and-fill method[15] does not make the overall pooled results become insignificant: 8.52 Nm (95% CI: 2.58, 14.46).

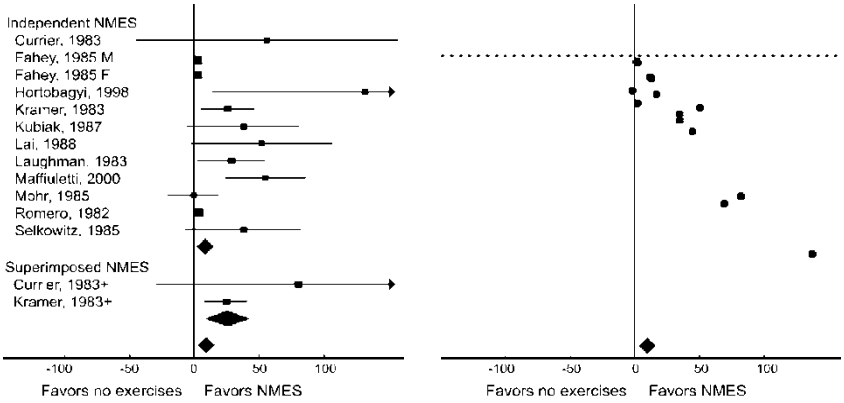


Figure 8.3 Plots for NMES versus no exercises – healthy quadriceps

A forest plot (left) and a funnel plot (right). The squares and circles represent the mean outcome of each study and the corresponding horizontal lines are 95% confidence intervals. The diamonds represent the pooled (subgroup) outcomes with the horizontal width corresponding to the outcome's 95% confidence interval. F =females; M = males.

One-way sensitivity analyses indicate that in both subgroups the pooled point estimates and their confidence intervals of the fixed effect model and the random effects model differ substantially: 3.57 (95% CI: 1.87, 5.28) versus 8.00 (95% CI: 2.79, 13.21) respectively. Nevertheless, although some differences are apparent, the point estimates and confidence intervals of both models are in favor of NMES. None of the trials used the common methods to reduce bias that were part of the sensitivity analysis factors. With regard to clinical parameters, usage of a volitional component in the stimulated contraction (superimposed NMES subgroup), pulse frequency, and pulse width stood out to some extent, with all factors positively (but not significantly) influencing the results of NMES.

8.4.3 NMES versus volitional exercises

The 'NMES versus volitional exercises' comparison group contains studies that investigated the effect of NMES as a substitute for volitional exercises. Again, a distinction is made between independent NMES and superimposed NMES. Of the ten studies eligible for this comparison group, two studies did not report the results in a manner that allowed usage of the data in meta-analysis[18,21]. The study by Hortobagyi et al. [21] favors NMES$^-$ based training, while the study by Singer et al. [18] found volitional isokinetic and isotonic exercises to be superior to NMES$^-$. Hortobagyi et al.[21] used a six-week eccentric training protocol, whereas Singer et al. [18] used a four-week concentric training protocol, which may account for some differences in results. Both trials may suffer from systematic error due to methodological issues (no allocation concealment, blinding, or intention-to-treat analysis). The results of the remaining studies[17,20,22,23,25,27,29,30] were included in the meta-analysis.

The weighted mean difference for the 'independent NMES versus volitional exercises' comparison is -11.60 Nm (95% CI: -24.34, 1.13) and is based on eight trial references and 155 subjects. Both the Chi-square value (χ^2=2.88; *df*=7; *P*=0.9) and the eyeball judgment of the overlap of the 95% confidence intervals of the trial results indicate only limited heterogeneity. The two studies that were eligible for the 'superimposed NMES versus volitional exercises' comparison favor volitional exercises, although the modes of the confidence interval functions (the point estimates) are positioned around the null value and include the null value. The weighted mean difference for this comparison is -11.11 Nm (95% CI: -37.06, 14.83) and is based on 37 subjects. The overall pooled point estimate is -11.51 Nm (95% CI: -22.94, -0.08).

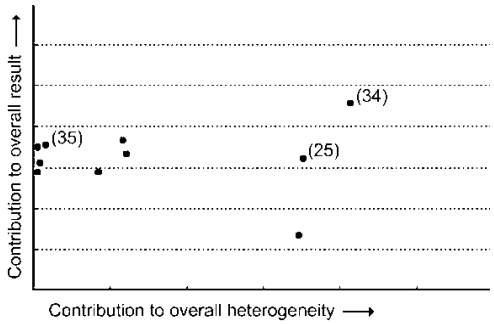


Figure 8.4 Heterogeneity of NMES versus volitional exercises – healthy quadriceps

The numbers in the figure correspond with the reference numbers used throughout the text.

In the heterogeneity exploration graph of both subgroups (Figure 8.4), two studies[20,29] had a relatively large influence on the overall result while also contributing substantially to the overall heterogeneity. The trial by Hortobagyi et al. [20] may stand out due to the measurement of the results in N instead of Nm and due to the eccentric training protocol. The other trial[29] has no exceptional clinical characteristics. Deleting them from the meta-analysis does not alter the results substantially, although the previous non-significant result of the 'independent NMES versus volitional exercises' comparison is now significantly in favor of volitional training. The trial by McMiken et al. [30], in which measurements were also in N instead of Nm, does not stand out in this heterogeneity analysis.

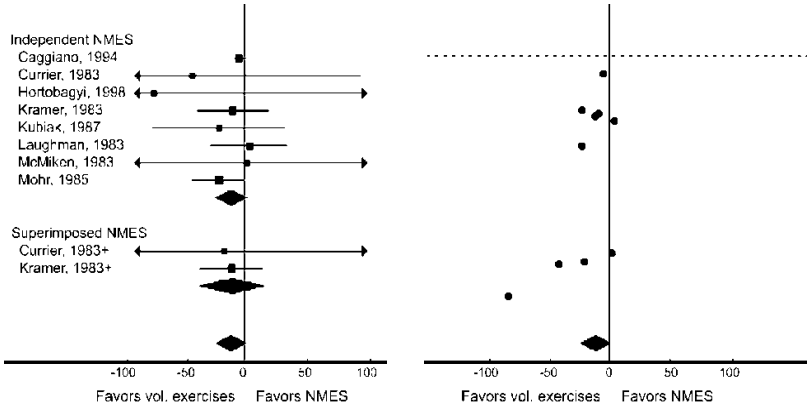


Figure 8.5 Plots for NMES versus volitional exercises – healthy quadriceps

A forest plot (left) and a funnel plot (right). The squares and circles represent the mean outcome of each study and the corresponding horizontal lines are 95% confidence intervals. The diamonds represent the pooled (subgroup) outcomes with the horizontal width corresponding to the outcome's 95% confidence interval. F =females; M = males.

The forest plot and funnel plot of both subgroups are presented in Figure 8.5. The funnel plot for this comparison group is also asymmetrical. The imputation of three trials via the trim-and-fill algorithm[15] makes the pooled point estimate shift towards the null value. The pooled point estimates of the fixed and random effects models are the same. This is in accordance with the limited statistical heterogeneity. In the sensitivity analyses, there was no indication of a substantial influence of deleting trials with the a-priori defined characteristics on the pooled point estimates.

8.4.4 Other comparisons

Independent versus superimposed NMES. Independent NMES was compared with superimposed NMES in the trials by Kramer and Semple[22] and Currier and Mann[25]. The point estimate of the latter study favors $NMES^{+}$, but the point estimate of the former study indicates no differences. Both trials have wide confidence intervals. The Delphi scores of the trials were 56%[22] and 44%[25], respectively, and neither of the two trials applied procedures such as allocation concealment or blinding. An additional methodological drawback of the trial by Currier and Mann is the large number of dropouts (19%) and the lack of exploration of its effect on the results.

NMES $\leq$ 40 Hz versus NMES $\geq$ 70 Hz. Only one trial, by Balogun et al. [16], was eligible for this comparison. In this small three-arm trial with ten subjects per arm, low frequency stimulation patterns of 80 Hz, 45 Hz, and 20 Hz were compared and the results consistently favor the NMES regime with the highest frequency (80 Hz). In the comparison between the 80 Hz pattern and the 20 Hz pattern, the confidence interval of the point estimate includes the null value and is in our calculations not significant. The high percentage of dropouts (27%) and the lack of additional analyses into the potential effect of this on the results seriously limit the interpretability of the results. The Delphi score of this study was 44%.

NMES in knee flexion versus NMES in knee extension. In the trial by Fahey et al. [26], the effect of NMES in 60 degrees knee flexion was compared with NMES in full knee extension. The men and women were randomized and analyzed separately. With regard to the main outcome of this review, no significant between-group differences (flexion and extension groups) could be found for either the male or the female group. Based on the full range of outcome measures, the authors find the results of the flexion-groups consistently (but not always significantly) better than those of the extension groups. Gender appeared to have no influence on the results. Like all other trials, this study lacks allocation concealment and blinding.

NMES at ≥ 50% MVC versus NMES at ≤ 30% MVC. Lai et al. [28] compared NMES in knee flexion at two intensities: 50% (or slightly more) and 25% (or slightly less) of the isometrically tested maximum volitional contraction of the quadriceps femoris. The authors used mean percentages of the change from baseline and describe significant ($p<0.05$, two-sided) between-group differences in favor of high intensity training. We used slightly different, more conservative calculations (with the absolute force measures, *z*-distribution and unequal variances assumption) and found a difference that was not significant: 30.3 Nm (95% CI: -25.09, 85.69). However, a large part of the confidence interval function lies above zero. The Delphi score of the trial was 56%.

8.5 Results – part 3: adults with impaired quadriceps femoris

Of the 18 included trials, two trials only had male participants[32,33] and the other 16 trials[34-49] had male and female subjects. Most trials used participants between 15 and 45 years of age, although two trials explicitly focused on elderly subjects[42,44]. Of the trials that could be included in a quantitative analysis, the average number of subjects per study arm was 16 (range 8-28). In addition to muscle torque, five studies[35,36,39,42,44] used outcome measures at the activity or participation level[13]. Most studies used isometric training methods with the knee flexed at 30 to 90 degrees angle and the average training period lasting five to six weeks. Various stimulation frequencies, pulse widths and training regimes were applied. Table 8.5 describes methods, study domain, interventions, and results of the included studies.

8.5.1 Quality assessment

The Delphi scores ranged from 11% to 67%, with a mean of 40% (SD: 13%). Thirty percent of the items of the Delphi list were answered with ‘don't know’. The Delphi scores (end scores as well as per-item scores) for each study are included in the third row of Table 8.5 and the order of the answers corresponds with the order of the Delphi list. Two studies used methods to conceal the randomization to the individuals responsible for the inclusion and exclusion of the participants[44,47]. Nine trials had no dropouts or loss to follow-up[32,33,35,37,38,40,41,46,49].

In some trials with a relatively large number of dropouts or loss to follow-up, e.g. larger than 10%[39,44,48], this type of analysis may have introduced some bias. Six trials used methods for blinded outcome assessment[37,39,42,43,44,47], but no evaluation of the failure or success of the blinding was attempted. In nine studies[32,34-36,38,42,44-47] the description of point estimates and dispersion of the relevant outcomes was insufficient for inclusion in the meta-analysis.

Table 8.5 Characteristics of the included studies with impaired subjects

ID and Subjects	Methods *	Bias †	NMES features
Anderson, 1989 Men and women (*n*=100), age 20-23, status after ACL rec.	5-arm RCT NMES$^-$ vs NO Other	Delphi score 22%. Y-?-Y-N-?-N-N-N-N	Isometric (?), no carrier, 35 Hz, 150 µs, hold 10 s, rest 110 s, for 60 minutes, 7 days a week, for 3 months. As adjunct.
Buhmann, 1998 Men and women (*n*=36), age 18-47, status after ACL rec.	3-arm RCT NMES$^-$ vs NO	Delphi score 11% Y-?-?-?-?-N-N-N-N	Isometric (?), no carrier, 50 Hz, hold 10s, rest 20 s, 30 minutes, twice a day, 7 days a week, 12 weeks. As adjunct.
Callaghan, 2001 Men and women (*n*=16), age 30, patello-femoral pain	2-arm RCT Other	Delphi score 44% Y-?-Y-Y-?-N-?-Y-N	1) Isometric, low frequency background, doublet of pulses in higher frequency burst, 200 µs, hold 10s, rest 50s). 2) Isometric, 35-45Hz, 350 µs, hold 10s, rest 50 s. 5xper day, 7x per week, 6 weeks
Delitto, 1988 Men (and women?), age 19-44 (*n*=20), after ACL rec.	2-arm RCT NMES$^+$ vs VOL	Delphi score 44% Y-?-?-N-Y-N-N-Y-Y .	Isometric, superimposed, quadriceps and hamstrings concurrently, 2500 Hz carrier, 50 Hz excitatory frequency, no ramp, hold 15 s, rest 50 s, 15 reps, 5 days a week, 3 weeks.
Draper, 1991 Men and women (*n*=30), age 18-41, after ACL rec.	2-arm RCT NMES$^+$ vs VOL Other	Delphi score 44% Y-?-?-Y-N-N-N-Y-Y	Dynamic, superimposed, no carrier, 35 Hz, ramp 4 s, hold 10 s, rest 20 s (including ramp down of 2 s), 10 reps, 2-5 sets, daily, 3 x per day, 6 weeks.
Gobelet, 1992 Men and women (*n*=120), age 13-63, patello-femoral synd	3-arm RCT NMES$^-$ vs VOL	Delphi score 56% Y-?-Y-Y-Y-N-N-Y-N	Isometric (?), no carrier, 50 Hz, 200 µs, hold 15 s, rest 45 s, 30 minutes, 2 x per day, 7 days per week, 4 weeks.
Gould, 1982 Men and women (*n*=30), age 18-25, cast from groin to toes for 2 weeks.	3-arm RCT NMES$^-$ vs NO NMES$^-$ vs VOL	Delphi score 56% Y-?-Y-Y-?-N-N-Y-Y	Isometric, no carrier, 37 Hz, 100 µs ramp-up 3 s, hold 5 s, rest 150 s, output at tolerance, 16 hours per day, 7 days per week, 2 weeks.
Halkjaer, 1985 Men (*n*=84), age?, medial coll. ligament injuries.	5-arm RCT NMES$^-$ vs NO NMES$^-$ vs VOL Other	Delphi score 33% Y-?-?-N-?-N-N-Y-Y	No carrier, (a) 10 Hz, (b) 50 Hz, 1 hour per day, 4 weeks.
Lieber, 1996 Men and women (*n*=40), age 15-44, after ACL surgery.	2-arm RCT NMES$^-$ vs VOL	Delphi score 44% Y-?-Y-Y-?-N-N-Y-?	Isometric, no carrier, 50 Hz, 250 µs, ramp up 2s, hold 8 s, rest 20 s, 30 min, 5x per week, 4 weeks.
Oldham, 1985 Elderly men and women with arthrosis (*n*=30), age 57-78.	4-arm RCT Other	Delphi score 44% Y-?-?-Y-Y-?-Y-N-?	Isometric (?), various frequency patterns, averaging 8.4 Hz, 300 µs, hold 30 s, rest 15 s, 3 hours, 7 days per week, 6 weeks.
Paternostro, 1999 Men and women (*n*=49), age 20-30, status after ACL surgery.	3-arm RCT NMES$^-$ vs NO Other	Delphi score 44% Y-?-?-Y-Y-N-N-Y-N Assessor blinding stated, not evaluated.	Isometric, 30-50 Hz, 200 µs, ramp-up 1-2 s, hold 5-10, rest 15-50, output maximally tolerable, 12 reps, 6 sets (4+2), 7x per week, 6 weeks. As adjunct.

* 'NMES$^-$', independent NMES; 'NMES$^+$', superimposed NMES; 'NO', no exercise group; 'VOL', volitional exercise. † Scores on Delphi items: 'Y', yes; 'N', no; '?', don't know.

Table 8.5 Characteristics of the included studies with impaired subjects (cont.)

ID and Subjects	Methods *	Bias †	NMES features
Quittan, 2001 Elderly men and women (*n*=42), mean age 58, with refractory heart failure,	2-arm RCT NMES^{-} vs NO	Delphi score 67% Y-Y-Y-Y-Y-N-N-Y-N Allocation concealed. Assessor blinding stated, not evaluated. Dropouts > 10%.	Isometric, 50 Hz, 700 µs, hold 2 s, rest 6 s, 25%-30% MVIC, 30-60 min, 5 days per week, 8 weeks.
Riel, 1990 Men (*n*=20), age 15-28, status after post-operative knee immob., wearing cast.	2-arm RCT NMES^{-} vs VOL	Delphi score 22% Y-?-?-?-?-N-N-Y-?	Isometric (within cast), 50 Hz, hold 5 s, rest 12.5 s, output powerful contraction (40-80 mA), 3x per day, 7 days per week, 3 weeks.
Sisk, 1987 Men and women, mean age 23 (*n*=24), status after ACL rec.	2-arm RCT NMES^{-} vs NO	Delphi score 33% Y-?-?-Y-?-N-N-Y-N	Isometric (within cast), 40 Hz, 300 µs (?), ramp-up o.5 s, hold 10 s, rest 30 s, output at visible contraction, 8 hours, 1x per day, 7x per week, 6 weeks. As adjunct.
Snyder-Mackler, 1991 Men and women (*n*=10), age 18-28, status after ACL rec.	2-arm RCT NMES^{+} vs NO	Delphi score 33% Y-?-?-Y-?-N-N-Y-? Analysis conform randomization.	Isometric, superimposed, 2500 Hz carrier, 400 µs, 75 Hz excitatory frequency, ramp-up 3 s, hold 12 s, rest 50 s, output at maximally tolerated level, 15 reps, 1 set, 3x per week, 4 weeks, As adjunct.
Snyder-Mackler, 1995 Men and women (*n*=129), age 15-43, status after ACL rec.	4-arm RCT NMES^{-} vs VOL Other	Delphi score 44% Y-Y-?-Y-Y-N-N-N-? Allocation concealed Assessor blinding stated, not evaluated.	(1) Isometric, 2500 Hz carrier, 75 Hz excitatory frequency, hold 11 s, rest 120 s, output maximally tolerable, 15 reps, 1 set, 3x per week, 6 weeks AND (2) 55 Hz excitatory frequency, 300 µs, hold 15 s, rest 50 s, output maximally tolerable, 15 minutes, 4x per day, 5x per week, 6 weeks
Wigerstad, 1988 Men and women, age 21-45 (*n*=26), status after ACL rec, wearing cast.	2-arm RCT NMES^{+} vs NO	Delphi score 44% Y-?-Y-Y-?-N-N-Y-N Dropouts > 10%, no additional analysis.	Isometric, superimposed, within-cast no carrier, 30 Hz, 300 µs, ramp-up 2 s, hold 6 s, rest 10 s, output maximally tolerable (60-100mA), 10 minutes, 4 sets (10 minutes rest), 3x per week, 6 weeks. As adjunct.
Williams, 1986 Men and women (*n*=21), age 18-45, status after menisect.	2-arm RCT NMES^{-} vs NO	Delphi score 33% Y-?-?-Y-?-N-N-Y-?	Isometric, 2500 Hz carrier, 50 Hz excitatory frequency, ramp-up 3.5 s, hold 11.5 s, rest 50 s, output at tolerance level, 10 reps, 1 set, 10 minutes, 5x per week, 3 weeks. As adjunct.

* 'NMES^{-}', independent NMES; 'NMES^{+}', superimposed NMES; 'NO', no exercise group; 'VOL', volitional exercise. † Scores on Delphi items: 'Y', yes; 'N', no; '?', don't know.

8.5.2 NMES versus no exercises

Besides comparing NMES-based training with no training at all, the 'NMES versus no exercises' comparison also includes studies that investigated the effect of NMES as an adjunct to basic exercise programs. In these latter studies all subjects were performing exercises and in the experimental study-arm, this was supplemented by NMES. Studies are divided into two subgroups: studies with

subjects that received stimulation or training after an immobilization period and studies with subjects that received stimulation during the immobilization period while wearing a cast or similar product. Formal meta-analysis was not feasible due to heterogeneity and reporting deficiencies. Figure 8.6 gives a 'metaview' of the studies in the subgroups, which can be used for visualizing the individual study outcomes.

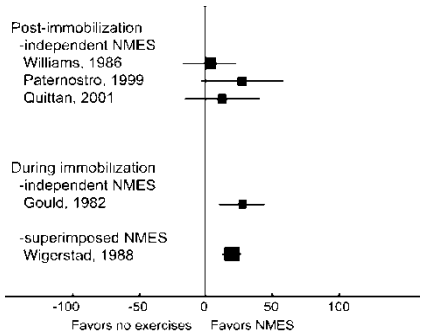


Figure 8.6. Metaview of NMES versus no exercises – impaired quadriceps.

The squares represent the mean outcome of each study and the corresponding horizontal lines are 95% confidence intervals.

Post-immobilization subgroup. Seven studies were eligible for this comparison[34,35,43,44,45,46,49]. Unfortunately, only three out of the seven studies reported the results in a format suitable for inclusion in a meta-analysis[43,44,49]. The results of these studies can be seen in the upper part of Figure 8.5. All of these studies show results that, to some extent, favor NMES. The study by Paternostro et al.[43] shows the most positive results, which may be due to the relatively high training and output intensity (seven times per week and output at maximally tolerable level). The least positive study by Williams et al.[49] did also use an intensive electrical output but the ten repetitions per training session for only three weeks may not have been of sufficient intensity to produce significant force increases. In the study by Quittan et al.[44], the subjects were all elderly men and women. Relatively low contraction intensities were used, but the extensiveness of the training program (eight weeks long, five times per week 30-60 minutes training) may have provided a sufficient training stimulus. This study also reported Quality-Of-Life and Activities-of-Daily-Living data, which were generally, but not always significantly, in favor of the NMES group. All studies in this comparison have substantial methodological weaknesses and although the study by Quittan et al. scores best on the Delphi list (67%), the high percentage of dropouts raises concern about the validity of the analysis.

Except for the study by Sisk et al.[44], all of the studies that did not provide the results in sufficient detail or format describe better results for NMES training with regard to strength augmentation[32,35,44,46]. One of the reasons for the relatively limited effect of NMES that was found in the study by Sisk et al. may be the fact that the current output of the NMES in this study was relatively low (the criterion for sufficient output was reported as "a visible contraction") and NMES was studied as an adjunct to basic exercises. The study by Snyder-Mackler et al.[46] used superimposed NMES as an adjunct to volitional exercises and more positive results are described for the NMES assisted training. Buhmann et al.[35] concluded that NMES should be recommended as an obligatory adjunct to treatment after anterior cruciate ligament surgery. Anderson and Lipscomb[32] reported significant improvement with regard to strength and amount of crepitation. All studies in this subgroup have substantial methodological or reporting drawbacks and their conclusions need to be interpreted with caution.

During-immobilization subgroup. Three studies were eligible to be included in this subgroup[32,40,48]. The generally used method in these trials was the application of electrodes on the skin through holes in the cast or after temporarily taking off the immobilizing brace. The results of the trials by Gould et al.[40] and Wigerstad et al.[48] were suitably reported for inclusion in an exploratory meta-analysis and are displayed in the lower portion of the metaview in Figure 8.6. Wigerstad et al. used superimposed NMES, whereas Gould et al. used independent NMES. Both trials show results that favor NMES, but the differences are not significant. This is also the case in the trial by Halkjaer-Kristensen et al.[32], although the data are not given in a format that allows inclusion in the meta-analysis. The trial by Gould et al. is exceptional to some extent, since the legs of healthy subjects, instead of post-operative subjects, were immobilized. Nevertheless, the subjects in all three trials displayed the same type of impairment due to immobilization and this impairment was, to some extent, prevented by NMES.

8.5.3 NMES versus volitional exercises

Post-immobilization subgroup. Five trials were suitable for this comparison[37,38,39,41,47], but only three trials[37,39,41] described the results in a manner that allowed inclusion in a meta-analysis. The latter are graphically displayed in the upper part of Figure 8.7, which gives a metaview of the individual trial outcomes without pooled point estimates. Of these studies, only the trial by Delitto et al.[37] yielded results that clearly favor NMES, which may be explained by the superimposition of NMES on volitional contractions, potentially resulting in a higher training stimulus compared to the volitional training group. In the trial by Gobelet et al.[39] one isometric volitional training arm and one isokinetic volitional

training arm were compared with NMES$^-$. Besides force outcomes, also Arpege knee-scores were documented. No significant differences were found between any of the trial-arms with regard to any of the outcomes, although all outcomes favored NMES$^-$ to some degree. All trials suffered from methodological weaknesses and although the study by Gobelet et al. scored best on the Delphi list (56%), the number of dropouts was substantial and no additional analysis into the possible consequences were performed.

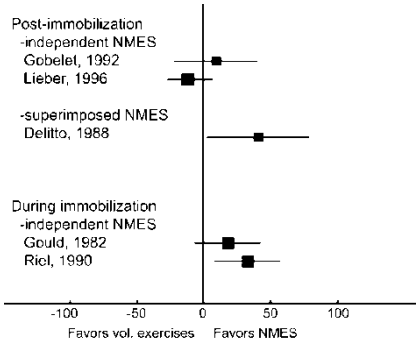


Figure 8.7 Metaview of NMES versus volitional exercises – impaired quadriceps

The squares represent the mean outcome of each study and the corresponding horizontal lines are 95% confidence intervals.

The remaining two trials[38,47] described their results only as relative percentages without sufficient data description to allow meta-analytic calculations. The trial by Draper and Ballard[38] shows results that favor (EMG-assisted) volitional training, whereas Snyder-Mackler et al.[47] conclude that only treatment with a medium frequency device has benefits over volitional exercises while treatment with a low-frequency portable NMES device does not. The latter conclusion seems rather immature since the stimulation regimes differed in many other respects than only the presence or absence of a carrier frequency. Both trials scored 44% on the Delphi list. The study by Snyder-Mackler et al.[47] used allocation concealment and states assessor blinding (not evaluated), but scores low on other Delphi items.

During-immobilization subgroup. Three studies were eligible to be included in this subgroup[32,33,40]. The results of the trials by Gould et al.[40] and Riel and Bernet[33] were appropriately reported for inclusion in a meta-analysis and are displayed in the lower portion of Figure 8.7. The trial by Gould et al. shows no significant between-group differences. In the trial by Riel and Bernet, on the other hand, significant between-group differences are present and favor NMES. The discrepancy between the results may be due to methodological issues (for which the trial by Gould et al. scores significantly better on the Delphi list: 56%

compared to 22%) or perhaps due to the difference in the population from which the sample was drawn. As mentioned earlier, the trial by Gould et al. stands out because healthy subjects, instead of post-operative subjects, were immobilized and trained to prevent immobilization-induced strength losses.

8.5.4 Other comparisons

Experimental NMES versus traditional NMES. Two trials compared traditional, constant frequency NMES with NMES based on variable frequency trains[36,42]. However, as far as the stimulation parameters are described, some questions can be raised about whether the training stimulus in both trials would have been sufficient to cause force augmentation. In the trial by Oldham et al.[42], the average stimulation frequency was only 8.5 Hz (generally frequencies between 30 Hz and 100 Hz are recommended and indeed used in most included trials) and in the trial by Callaghan et al.[36] the details of the actual stimulation pattern were not reported at all. Considering the clinical heterogeneity, comparison of the two studies is impossible. Both studies report various outcomes, such as force augmentation, sit-to-stand time, number of step-ups, and timed ten-meter walk, but no significant differences (alpha=0.05, double-sided) were found.

8.6 Discussion

The clinical applicability of systematic reviews of randomized controlled trials depends heavily on the availability and methodological quality of the studies that are eligible for inclusion. With regard to the availability of trials for this review, the question arises whether the published and retrieved studies represent the full spectrum of trials performed in actual research practice. The funnel plot in the meta-analytic data syntheses suggests that publication bias is likely to be present. Fortunately, imputing 'missing trials' by means of the trim-and-fill method does not lead to different conclusions than those of the original analysis, illustrating that the general conclusions of the meta-analyses may be relatively robust with regard to publication bias.

Although we used change from baseline scores wherever possible, we have used standard deviations of follow-up scores for all studies in the meta-analyses to make sure the studies were weighted under equal conditions. Although using standard deviations of change scores would have provided more power, several of the studies that did provide data with regard to the mean follow-up scores and their variance did not provide sufficient data to extract or calculate the mean change from baseline scores or their variance estimates. Estimated variance imputations for change from baseline scores can be an option at times[32], but since a relatively large number of studies did not provide sufficient data in this respect,

we decided to consequently use the variance estimates of the follow-up scores. This resulted in weighting under equal conditions, but made the meta-analysis somewhat more conservative than when the substantially smaller variances of the actual mean changes from baseline could have been used.

As for the methodological quality, the abundance of methodological and descriptive weaknesses in the included studies poses a serious threat to the assumption that randomized controlled trials essentially have the potential to provide sound evidence on the effects of health care interventions. Although a randomization procedure was performed in each of the trials, the lack of simple bias-limiting methods such as allocation concealment, assessor or statistician blinding, and intention-to-treat analyses meant that various forms of bias could not be ruled out. Detailed and complete description of the procedures, the intervention characteristics, and the results is also lacking in many trials, while this should actually be a prerequisite for trials to get published.

We did not contact the authors for information about data that were not reported in the trials, since non-response is often high and data-reliability doubtful. As the missing data were mostly related to methodological issues and intervention characteristics and only to a small extent to the description of point estimates[18,19,21], additional information from the authors would likely have had little impact on the quantitative results of review.

We assessed the validity aspects of the trials via per-item and summary scores of the Delphi list[12]. Although we are aware of the difficulties of summary scores as a meta-analytic tool[51-53], we felt that these scores could be useful for clinicians as general indicators of the potential value of the study. We did not use the scores as cut-off points for analyses.

The methodology assessment with the Delphi list and the data extraction were not performed in a blinded manner. As the usefulness of blinding in systematic reviews is still subject of discussion[54-56], we decided not to change the article format, but instead we used an extra assessor (three assessors instead of two) and asked three additional reviewers (total of six reviewers) to join the consensus sessions. We expect these procedures to have minimized the bias that could be associated with journal recognition, author recognition, and/or knowledge of the results.

Some may argue that data are clinically too heterogeneous to perform a meta-analysis, and in fact we did consider using only a descriptive approach without

data pooling. Nevertheless, since the heterogeneity in two of our comparisons was not involving the direction of the effect but rather its magnitude (with minimal clinical relevance of this magnitude) we found a quantitative approach justified for these comparisons. Another consideration was that the impact of changing various parameters might, considering our data, in fact be very small. Our first option was a meta-regression approach, but the number of potentially relevant parameters and other clinical factors was too large in relation to the number of trials and participants. Apart from making us reflect on our own decisions, the heterogeneity issue clearly underlines the need for researchers in the field of clinically applicable research to carefully assess the domain details and intervention characteristics described in the trials included in this review (for details see Tables 8.4 and 8.5), choose clinically applicable, commonly used and validated outcome measures, and make their research approach as standardized as possible.

A further potential limitation of the review is the fact that the main outcome measure is, according to the ICF definitions[13], at a structural/functional level and not at an activity or participation level. The choice of muscle torque as main outcome measure is largely based on the limited availability of trials that have used outcome measures on an activity or participation level. However, based on the limited increases in muscle torque found in this review, most questionnaires and scales designed to evaluate musculoskeletal activity status may be too insensitive to measure the gain in activity performance of healthy individuals. In this respect specialized activity tests may be useful for future trials.

The available evidence is of limited quality and needs to be interpreted with caution. However, most of the evidence indicates that NMES can be an effective modality to increase quadriceps muscle strength. Especially in cases where patients are wearing a cast, NMES may even be more effective than volitional training for minimizing the strength loss in the immobilization period. In all other cases the evidence indicates that volitional exercises may be equally or more effective. The presence of a volitional component in the NMES-induced contraction appears relevant for the efficacy of NMES, but additional research is necessary to further quantify optimal parameter settings. Based on the current evidence, it can be concluded that for healthy subjects and post-immobilization patients, NMES is likely to be more appropriate as an adjunct to, rather than a replacement of, volitional (quadriceps) strength training. For patients wearing casts, applying NMES via holes in the cast may be valuable to decrease the loss of strength. Data about the impact of the strength improvements on activity or participation levels are scarce and all conclusions should be regarded as preliminary until additional randomized controlled trials of better quality have

been performed, although bias-limiting strategies will need to be implemented more rigorously than has been done so far. Finally, when reporting results of clinical research, researchers should realize that their primary audience consists of clinicians and that a detailed and complete description of the procedures, the intervention characteristics, and the results is absolutely essential for the trial to have a legitimate impact on clinical practice.

Acknowledgments

The research was not supported by any grants. The authors acknowledge Katrien Bartholomeeusen, Simon Brumagne, Sara van Deun, and Koen Janssens for their assistance during the validity assessments.

Conflict of interest

The first author has received financial compensation for consulting services from ITO Co. Ltd., a manufacturer of electrotherapy equipment.

References

1. Stillings D. Electrical stimulation for foot drop, 1772. *Med Instrum* 1975 Nov-Dec; 9 (6): 276-7.

2. Pichon F, Chatard JC, Martin A, et al. Electrical stimulation and swimming performance. *Med Sci Sports Exerc* 1995; 27 (12): 1671-6.

3. Maffiuletti NA, Cometti G, Amiridis IG, et al. The effects of electromyostimulation training and basketball practice on muscle strength and jumping ability. *Int J Sports Med* 2000; 21: 437-43.

4. Harbour R, Miller J. A new system for grading recommendations in evidence based guidelines. *BMJ* 2001; 323: 334-6.

5. Lloyd T, De Domenico G, Strauss GR, et al. A review of the use of electro-motor stimulation in human muscles. *Aust J Physiother* 1988; 32 (1): 18-30.

6. Callaghan MJ, Oldham JA. A critical review of electrical stimulation of the quadriceps muscles. *Crit Rev Phys Rehab Med* 1997; 9 (384): 301-14.

7. Hainaut K, Duchateau J. Neuromuscular electrical stimulation and voluntary exercise. *Sports Med* 1992; 14 (2): 100-13.

8. Kramer JF, Mendryk SW. Electrical stimulation as a strength improvement technique: a review. *J Orthop Sports Phys Ther* 1982; 4 (2): 91-8.

9. Selkowitz DM. High frequency electrical stimulation in muscle strengthening. *Am J Sports Med* 1989; 17 (1): 103-11.

10. Marks R, Ungar M, Ghasemmi M. Electrical stimulation for osteoarthritis of the knee: biological basis and systematic review. *N Z J Physiother* 2000; 28 (3): 6-21.

11. Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. *BMJ* 1994; 309: 1286-91.

12. Verhagen A.P., Vet HCW de, Bie RA de, et al. The Delphi List: A Criteria List for Quality Assessment of Randomized Clinical Trials for Conducting Systematic Reviews Developed by Delphi Consensus. *J Clin Epidemiol* 1998; 51 (12): 1235-41.

13. World Health Organization (WHO), 2001. International Classification of Functioning, Disability and Health (ICF). Geneva, World Health Organization, 2001.

14. Baujat B, Mahé C, Pignon JP, et al. A graphical method for exploring heterogeneity in meta-analyses: application to a meta-analysis of 65 trials. *Stat Med* 2002; 21: 2641-52.

15. Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* 2000; 56 (2): 455-63.

16. Balogun JA, Onilari AO, Akeju OA, et al. High voltage electrical stimulation in the augmentation of muscle strength. *Arch Phys Med Rehabil* 1993; 74: 910-6.

17. Caggiano E, Emrey T, Shirley S, et al. Effects of electrical stimulation or voluntary contraction for strengthening the quadriceps femoris muscles in an aged male population. *J Orthop Sports Phys Ther* 1994; 20 (1): 22-8.

18. Singer KP, Gow PJ, Otway WF, et al. A comparison of electrical muscle stimulation, isometric, isotonic and isokinetic strength training programs. *N Z J Sports Med* 1983; 11 (3): 61-3.

19. Wolf SL, Ariel GB, Saar D, et al. The effect of muscle stimulation during resistive training on performance parameters. *Am J Sports Med* 1986; 14 (1): 18-23.

20. Hortobagyi T, Lambert J, Scot K. Incomplete muscle activation after training with electromyostimulation. *Can J Appl Physiol* 1998; 23 (3): 261-70.

21. Hortobagyi T, Scot K, Lambert J, Hamilton G, et al. Cross-education of muscle strength is greater with stimulated than voluntary contractions. *Motor Control* 1999; 3: 205-19.

22. Kramer JF, Semple JE. Comparison of selected strengthening techniques for normal quadriceps. *Physiother Can* 1983; 35 (6): 300-4.

23. Mohr T, Carlson B, Sulentic C, et al. Comparison of isometric exercise and high volt galvanic stimulation on quadriceps femoris muscle strength. *Phys Ther* 1985; 65 (5): 606-9.

24. Romero JA, Sanford TL, Schroeder RV, et al. The effects of electrical stimulation of normal quadriceps on strength and girth. *Med Sci Sports Exerc* 1982; 14 (3): 194-7.

25. Currier DP, Mann R. Muscular strength development by electrical stimulation in healthy individuals. *Phys Ther* 1983; 63 (6): 915-921.

26. Fahey TD, Harvey M, Schroeder RV, et al. Influence of sex differences and knee joint position on electrical stimulation modulated strength increases. *Med Sci Sports Exerc* 1985; 17 (1): 144-7.

27. Kubiak RJ, Whitman KM, Johnston RM. Changes in quadriceps femoris muscle strength using isometric exercise versus electrical stimulation. *J Orthop Sports Phys Ther* 1987; 8 (11): 537-41.

28. Lai HS, De Domenico G, Strauss GR. The effect of different electro-motor stimulation training intensities on strength improvement. *Aust J Physiother* 1988; 34 (3): 151-64.

29. Laughman RK, Youdas JW, Garret TR, et al. Strength changes in the normal quadriceps femoris muscle as a result of electrical muscle stimulation. *Phys Ther* 1983; 63 (4): 494-9.

30. McMiken DF, Todd-Smith M, Thompson C. Strengthening of human quadriceps muscles by cutaneous electrical stimulation. *Scand J Rehabil Med* 1983; 15: 25-8.

31. Selkowitz DM. Improvement in isometric strength of the quadriceps femoris muscle after training with electrical stimulation. Phys Ther 1985; 65 (2): 186-96.

32. Halkjaer-Kristensen J, Ingemann-Hansen T. Wasting of the human quadriceps muscle after knee ligament injuries. *Scand J Rehab Med* 1985; Supplement 13: 29-37.

33. Riel KA, Bernet P. Transkutane elektrische Muskelstimulation wahrend der postoperativen immobilisation des Kniegelenkes [Transcutaneous electrical muscle stimulation during post-operative knee immobilization]. *Sportmedizin* 1990; 41: 425-8.

34. Anderson AF, Lipscomb AB. Analysis of rehabilitation techniques after anterior cruciate reconstruction. *Am J Sports Med* 1989; 17 (2): 154-60.

35. Buhmann HW, Schleicher W, Urbach D, Schulz W. Elektromyostimulation und isokinetisches Training in der Rehabilitation nach Operationen des vorderen Keuzbandes - eine randomisierte, prospective Studie [Electromyostimulation and isokinetic training in rehabilitation after anterior cruciate surgery]. *Physik Medizin, Rehabilitationsmedizin, Kurortmedizin* 1998; 8 (1): 13-6.

36. Callaghan MJ, Oldham JA, Winstanley J. A comparison of two types of electrical stimulation of the quadriceps in the treatment of patellofemoral pain syndrome. *Clin Rehab* 2001; 15: 637-46.

37. Delitto A, Rose SJ, KcKowen JM, Lehman RC, Thomas JA, Shively RA. Electrical stimulation versus voluntary exercise in strengthening thigh

musculature after anterior cruciate ligament surgery. *Physical Therapy* 1988; 68 (5): 660-3.

38. Draper V, Ballard L. Electrical stimulation versus electromyographic biofeedback in the recovery of quadriceps femoris muscle function following anterior cruciate ligament surgery. *Physical Therapy* 1991; 71 (6): 455-64.

39. Gobelet C, Frey M, Bonard A. Techniques de musculation et chondropathie retro-patellaire [Muscle development techniques and retropatellar chondropathy]. *Rev Rhum* 1992; 59 (1): 23-7.

40. Gould N, Donnermeyer D, Pope M, Ashikaga T. TRanscutaneous muscle stimulation as a method to retard disuse atrophy. *Clin Orthop Rel Res* 1982; 164: 215-20.

41. Lieber RL, Silva PD, Daniel DM. Equal effectiveness of electrical and volitional strength training for quadriceps femoris muscles after anterior cruciate ligament surgery. *J Orthop Res* 1996; 14 (1): 131-8.

42. Oldham JA, Howe TE, Petterson T, Smith GP, Tallis RC. Electrotherapeutic rehabilitation of the quadriceps in elderly osteoarthritic patients: a double blind assessment of patterned neuromuscular stimulation. *Clin Rehab* 1995; 9 (1): 10-20.

43. Paternostro-Sluga T, Fialka C, Alacamliogliu Y, Saradeth T, Fialka-Moser V. Neuromuscular electrical stimulation after anterior cruciate ligament surgery. *Clin Orthop Rel Res* 1999; 368: 166-75.

44. Quittan M, Wiesinger GF, Sturm B, Puig S, Mayr W, Sochor A, Paternostro T, Resch KL, Pacher R, Fialka-Moser V. Improvement of thigh muscles by neuromuscular electrical stimulation in patients with refractory heart failure. *Am J Phys Med Rehab* 2001; 80 (3): 206-14.

45. Sisk TD, Stralka SW, Deering MB, Griffin JW. Effect of electrical stimulation on quadriceps strength after reconstructive surgery of the anterior cruciate ligament. *Am J Sports Med* 1987; 15 (3): 215-20.

46. Snyder-Mackler L, Ladin Z, Schepsis AA, Young JC,. Electrical stimulation of the thigh muscles and reconstruction of the anterior cruciate ligament. *J Bone Joint Surgery* 1991; 73-A (7): 1025-36.

47. Snyder-Mackler L, Delitto A, Bailey SL, Stralka SW. Strength of the quadriceps femoris muscle and functional recovery after reconstruction of the anterior cruciate ligament. *J Bone Joint Surgery* 1995; 77-A (8): 1166-73.

48. Wigerstad-Lossing I, Grimby G, Jonsson T, Morelli B, Peterson L, Renstrom P. Effects of electrical muscle stimulation combined with voluntary contractions after knee ligament surgery. *Med Sc Sports Exerc* 1988; 20 (1): 93-8.

49. Williams RA, Morrissey MC, Brewster CE. The effect of electrical stimulation on quadriceps strength and thigh circumference in menisectomy

patients. *J Orthop Sports Phys Ther* 1986; 8 (3): 143-6.

50. Follmann D, Elliot P, Suh I, et al. Variance imputation for overviews of clinical trials with continuous response. *J Clin Epidemiol* 1992; 45 (7): 769-73.

51. Huwiler-Muntener K, Jüni P, Junker C, et al. Quality of reporting of randomized trials as a measure of methodological quality. *JAMA* 2002 Jun 5; 287 (21): 2801-4.

52. Juni P, Witschi A, Bloch R, et al. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA. 1999 Sep 15; 282 (11): 1054-60.

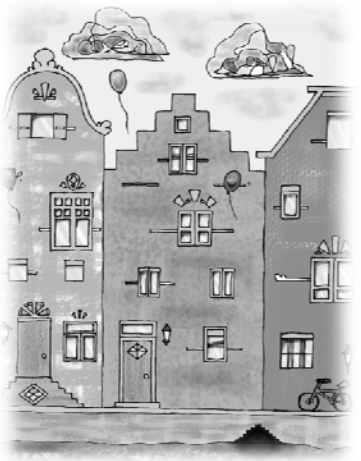
53. Greenland S. Can meta-analysis be salvaged? *Am J Epidemiol* 1994 Nov 1; 140 (9): 783-7.

54. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials* 1996; 17: 1-12.

55. Berlin JA. Does blinding of readers affect the results of meta-analyses? *Lancet* 1997; 350: 185-6.

56. Verhagen AP, Vet HCW de, Bie RA de, et al. Balneotherapy and quality assessment; inter-observer reliability of the Maastricht criteria list and the need for blinded quality assessment. *J Clin Epidemiol* 1998; 51: 335-41.

Discussion and Summary



9

META-ANALYSIS BEYOND META-SYNTHESIS

Based on:
Bax L, Diamond G, Kaul S, Moons KG.
Meta-analysis beyond meta-synthesis.
Submitted for publication.

Abstract

Introduction. Meta-analysis is a popular tool for summarizing medical evidence. Commonly, the primary objective is to create a single summary estimate from a set of studies.

Methods. We plead for a structurally different approach where a meta-analysis is divided in stages and the summary or synthesis is only a small part of the meta-analytical process.

Results. The paper gives a practical roadmap of the exploration, synthesis, and evaluation processes in causal meta-analysis. Instructions and syntax for the free statistical software packages MIX, R, and OpenBUGS are provided to facilitate the application of the proposed methods.

Conclusion. Meta-analysis is more than a method to establish a (significant) pooled result based on aggregate-level data from multiple studies. Meta-analysts should routinely explore and evaluate sensitivity of results to alternative statistical approaches and provide the reader with evidence for the robustness of the presented results.

9.1 Introduction

Meta-analysis should be seen as the quantitative part of a formal study of previously conducted research (systematic review)[1-3]. The number of reviews of medical research that contain some form of meta-analysis has increased exponentially over the last decade[4,5]. They are increasingly used to guide clinical and health policies. Consequently, meta-analysts have a tremendous responsibility to provide results that truly represent the underlying evidence. Comprehensive meta-analysis is therefore not a trivial task. Substantial mathematical and statistical procedures are required to comply with the demand for valid scientific evidence. As Poincaré pointed out almost a century ago, a scientific exercise is more than an accumulation of facts[6].

A number of overviews of contemporary methods in meta-analysis of medical research data have been published in medical journals. Some are statistically oriented and provide insight in the methodological advances and mathematical backgrounds of meta-analysis[4,7]. Others take a more didactical approach, often with less attention to statistical detail[5,8,9]. Rather than create a state-of-the-art overview of contemporary methods for meta-analysis, we have attempted to restructure the methods in a practical framework. The framework is based on a distinction between data exploration, synthesis, and evaluation, and its application is illustrated with practical examples. To serve future researchers involved in meta-analysis, instructions and code are provided to recreate the output in the examples with the free meta-analysis software MIX, R, and OpenBUGS. The paper starts with a short overview of meta-analysis.

9.2 Meta-analysis - conceptions and misconceptions

9.2.1 Meta-analysis

The words ‘meta’ and ‘analysis’ have their etymological origins in Ancient Greek. Meta (μετα) meant ‘after’ or ‘beyond’, and analysis comes from the verb αναλυω, which translates loosely as ‘to unravel’ or ‘take apart’. Hence, meta-analysis is an analysis that comes after or incorporates other previous analyses. Traditionally, the units of analysis have been studies – so-called ‘aggregate-level analyses’. Increasingly, however, meta-analyses use primary patient data and incorporate analyses at the ‘individual level’. The latter are commonly referred to as individual patient data (IPD) meta-analyses.

Because data from different sources are combined, meta-analyses have also been referred to as ‘pooled analyses’ or ‘data syntheses’. The term ‘synthesis’ also comes from Ancient Greek (συνθεσις) and means ‘integration’ or ‘combination’,

which is in fact the antonym of analysis. This contrast, we believe, reveals the two major objectives a meta-analysis should have: (1) to simplify the interpretation of a multitude of compatible data from separate primary studies by numerical 'synthesis', and (2) to explore and clarify the potential causes of inevitable differences between the studies by an evaluative 'analysis'.

9.2.2 Critiques

Newness

There has been a fierce debate about the place of meta-analysis in medical science. Some argued that re-analyses of data from published research reports could not bring any 'neuess', a critique that is still heard today[10]. A common notion was that 'true' scientists should collect data from the primary units of interest (in medicine commonly human beings), and not from other scientists' reports.

Apples, oranges, and fruit

A common methodological objection is that studies with different methods, different study populations, and different treatment protocols should not be combined to get an overall estimate of the association that is studied; the famous argument not to mix 'apples with oranges' or one will end up with a fruit salad[11]. A related argument is that low-quality primary studies will also lead to low-quality meta-analyses. To stay with the fruit analogy, one should not mix good apples with bad apples to learn about the good ones. Another commonly heard expression in this context is 'garbage in, garbage out'[12].

The possible effect of selective processes in the dissemination of evidence (resulting in dissemination bias) is cited as another reason to refrain from meta-analysis. When it is unknown how much of the evidence has been suppressed (how many apples are missing and whether the remaining ones are representative of the whole), it could be argued that one should indeed refrain from a meta-analysis[13].

Pitfalls of the aggregate

Statistically, meta-analysis has been referred to as 'mega-silliness'[14], 'statistical alchemy for the 21st century'[15], and a 'statistical nightmare'[16].The primary objection of many statisticians to meta-analysis was that using the aggregate to infer about the individual inevitably leads to inferential uncertainty (and potentially bias) that is commonly not modeled in the meta-analysis process[16]. Additionally, aggregate-level meta-analyses are unable to account for individual-level data trends. This can result in aggregation bias, also referred to as the ecological fallacy[17,18] (Figure 9.1).

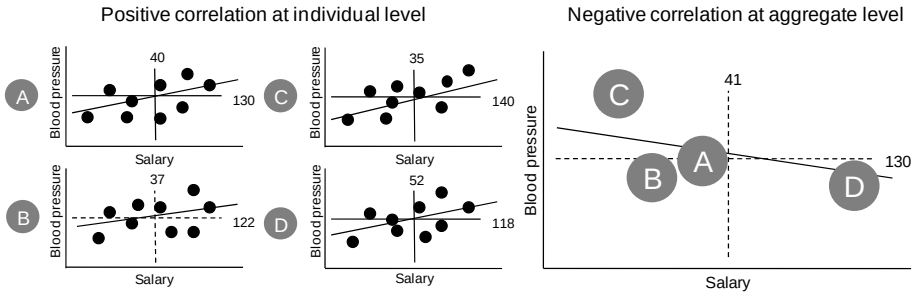


Figure 9.1 The ecological fallacy

The left part shows four studies (A-D) in which there is a negative correlation between salary and blood pressure. The right part shows how a different (in this case opposite) correlation can occur if the studies are analyzed at the aggregate level with only the means of blood pressure and salary. This phenomenon is called an ecological fallacy and the bias is commonly referred to as aggregation bias.

9.2.3 Critiques rebutted

Although the critiques are justified, they are often rooted in a narrow perspective of meta-analysis. We will address some of these critiques and propose an alternative view where appropriate.

Newness

Even though meta-analysis is by definition a tool that uses data that are not new and perhaps already published by others, it is a misconception that meta-analysis and its inferences cannot bring new and valuable insights. One of the important and unique contributions of meta-analysis is that it allows the analyst to evaluate whether existing evidence is (in)conclusive enough to justify exposing patients to potentially harmful situations in order to produce new evidence. This insight in itself is essential for the efficiency and ethics of science. If meta-analysis, perhaps combined with formal decision analysis, indicates that no single study could alter the significance or clinical relevance of a meta-analytical summary estimate, no additional study would be meaningful. Performing one nonetheless is arguably unethical. On the other hand, we would like to stress that if additional evidence is necessary, studies with identical research questions, hypotheses, design, data collection, and analysis (but obviously different samples) are essential for science to become re-search: it is in fact the results of such studies that are ultimately suitable to be explored in a meta-analysis.

Apples, oranges, and fruit

The debates about whether one should combine apples and oranges, bad apples with good apples, and apples from which some are missing are based on the

assumption that the primary objective of meta-analysis is the synthesis (of apples). If the objective is indeed solely to make a synthesis of information from different studies, it may be questionable whether the combined result of certain studies makes sense for the target population. However, if one is to go beyond the synthesis and the aim is to explore and analyze the reasons for the differences or the extent and impact of the missingness, meta-analytical techniques are indispensable. Just as it can make sense to generalize among fruit, it can also make sense to generalize among studies that have studied samples of different populations.

Pitfalls of the aggregate

With regard to the ecological fallacy and investigating covariates, we believe it is important to make a clear distinction between aggregate-level (study-level) covariates and individual-level covariates. Aggregate-level covariates such as study region or design characteristics are singular items per study and can be investigated without causing ecological or aggregation bias. We discourage the inclusion of means of individual-level covariates, e.g. blood pressure, education level, age, in an aggregate-level framework, because the association of the means with the outcome may not be the same as the association of the individual values and the outcome (Figure 9.1). This bias can be avoided by using the individual-level data in a (hierarchical) meta-analysis[19,20].

9.2.4 A framework for meta-analysis

As we noted in section 9.2.1, the two major objectives of a meta-analysis should be the synthesis and analysis of the data. In the synthesis a summary is made that facilitates practical inferences from the data set. In our view, a good synthesis is preceded and followed by analytical parts. The first is an exploration of the characteristics of the raw data that are to be used in the synthesis. The part following the synthesis attempts to clarify the reasons for differences between the studies and investigates the robustness of the synthesis's results to changes in statistical methods. We therefore propose to perform meta-analysis in a temporal framework of exploration, synthesis, and evaluation (Figure 9.2).

Figure 9.2 A framework for meta-analysis

9.3 Notation and data

9.3.1 Notation

Meta-analytical research uses hierarchical data from participants i (from 1 to n) nested in a number of scientific studies j (from 1 to k). The i units of analysis are often referred to as individual participant level data and the j units of analysis are aggregate-level data. Traditional meta-analysis uses only aggregate-level data that summarize each study by a single association measure y_j (e.g. a (log-transformed) odds ratio or risk ratio) and a measure of its inferential uncertainty (a variance, standard error, or confidence interval). Statistics like the odds ratio are calculated from the data and used to infer about the underlying (true) parameter of that observed effect. This parameter is commonly depicted by the Greek letter theta θ (Figure 9.3).

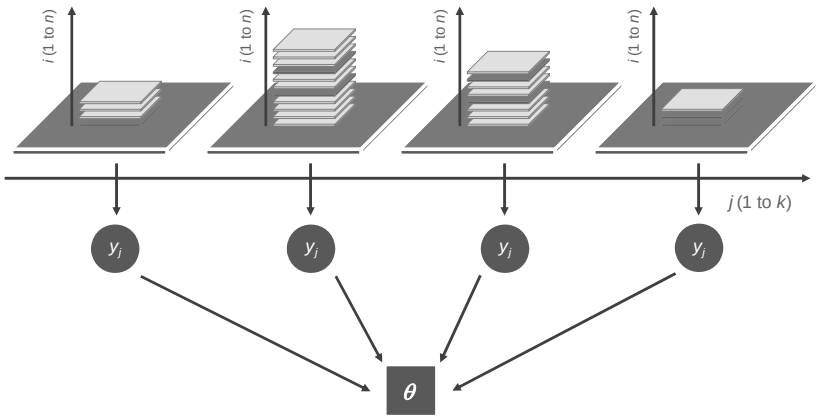


Figure 9.3 Estimation in the hierarchical structure of meta-analysis data

The vertically stacked small plates indicate participants i, which are nested in the horizontally aligned larger plates that symbolize studies j. Each study produces an estimate y_j that can be used for statistical inference of the underlying parameter θ (in this case a single, fixed effect). The arrows indicate the direction of the inference, not causality.

9.3.2 Data and measures of association

Most studies that are included in causal meta-analyses involve an index group (exposed to treatment or other potentially causal factor) and a reference group (not exposed to the index treatment or factor). Of interest is mostly the comparison between these groups, which is reported by a measure of association such as a risk ratio or odds ratio for binary outcomes, or (standardized) mean differences or correlation coefficients for continuous outcomes. These measures of association, together with a measure of the within-study variance, are the primary building blocks of many aggregate-level meta-analyses. In many cases, however, not only

are such comparative data available, but also the events and non-events that were used to calculate these measures for each group.

Throughout this paper, we will use a data set from a recently published systematic review and meta-analysis[21] to illustrate the methods that are discussed. The review assessed the adverse effects of perioperative beta-blockers in patients undergoing non-cardiac surgery. We chose the data from the subgroup in which the effect of the beta-blockers on non-fatal myocardial infarction was assessed. Table 9.1 shows data of the 22 included trials. The table has seven columns with the study identifier (id), event rate in the beta-blocker group (g1.r1), non-event rate in the beta-blocker group (g1.r0), number of subjects in the beta-blocker group (g1.n), event rate in the control group (g0.r1), non-event rate in the control group (g1.r0), and the number of subjects in the control group (g0.n).

Table 9.1 The beta-blocker data set

Study identifier	Events group 1	No events in group 1	Subjects group 1	Events group 0	No events in group 0	Subjects group 0
id	**g1.r1**	**g1.r0**	**g1.n**	**g0.r1**	**g0.r0**	**g0.n**
Jacobsen I	1	17	18	0	18	18
Lai	3	27	30	3	27	30
Liu	2	13	15	5	10	15
Magnusson	0	19	19	1	20	21
POBBLE	1	54	55	4	44	48
Poldermans	0	59	59	9	44	53
Urban	1	51	52	3	52	55
Zaug	0	43	43	3	17	20
Jacobsen II	0	50	50	0	50	50
Stone	0	89	89	0	39	39
DIPOM	3	459	462	2	457	459
MaVS	19	227	246	19	231	250
POISE	141	4033	4174	215	3962	4177
Raby	0	15	15	1	10	11
Rosenburg	0	19	19	1	18	19
Wallace	1	98	99	2	99	101
BBSA	0	112	112	0	112	112
Bayliff	0	49	49	0	50	50
Cuccharia	0	37	37	0	37	37
Davies	0	20	20	0	20	20
Jacobsen III	0	9	9	0	10	10
Miller	0	368	368	0	180	180

The data set is provided in CSV file format as Web appendices. In MIX, the file can be opened via **'Data wizard'** > **'Open CSV file'**, after which the columns

need to be selected to create a data set. Appendix B.1 (load data.r) contains the code to load the data set in R.

9.4 Exploration

9.4.1 Raw data

Before any statistical procedure is applied, one should inspect the raw data. This provides the analyst with the opportunity to inventory unlikely values, outliers, and missingness[22]. In MIX, this is done via the **'Exploration' > 'Data set summary'** menu. The R code to get similar information is provided in Appendix B.2 (data set exploration.r).

In the beta-blocker data set there are 22 studies, with a total number of 11,815 subjects. The mean and median of the sample sizes in the included studies are 537 and 100, respectively, indicating a skewed distribution and possibly the presence of outliers. Closer inspection of the data table shows that the POISE study (8351 subjects) and Miller study (548 subjects) are relatively big compared to the other studies. The smallest study (Jacobsen III) has 19 subjects.

The medians of the number of subjects in the treatment and control arms are 50 and 44. The medians of the number of events in the arms are 0 and 1. Thus, myocardial infarction is a rare event in the sample of subjects in this study. There are no missing values, but we see that six studies have one study arm in which no events occurred and eight studies have reported no events at all. The latter has implications for the calculation of ratio measures of effect size, such as the risk ratio and the odds ratio. For example, the risk ratio for the first study (Jacobsen I) cannot be calculated if the data are used as is. This can be solved by so-called continuity corrections, e.g. adding 0.5 to all event values of a study with a zero-event. Studies with no events at all are commonly excluded. There are a number of alternatives to these methods and we will elaborate on this in the analysis section.

9.4.2 Modeling assumptions

In Figure 9.3 it can be seen how the synthesis of estimates y_j from a number of studies can provide inference for the parameter θ. The real world works the other way around and the data, in this case the (log) odds ratios in the included studies, are 'caused' by the parameter(s)[6,23]. Viewing Figure 9.3 from this perspective would mean that the arrows should be reversed[23]. Statistical models assume that the values found in the data set are based on certain probability distributions. It is the challenge for the analyst to find a model, verify that it fits the data, and validate the proposition that it can be generalized beyond the data set.

If we assume, as depicted in Figure 9.3, that there is only one θ underlying the estimate from each of the included studies, we can speak of a singular parameter or a so-called 'fixed effect assumption'. This means that the distribution of the estimates y_j are assumed to come from a normal distribution with mean θ and variance σ^2:

$$y_j \sim N(\theta, \sigma^2) \,. \tag{1}$$

It should be noted that for ratio measures like the risk ratio or odds ratio, a log-transform is commonly applied before the effect sizes are modeled. In a fixed effect meta-analysis, we calculate a summary estimate that we believe is the best representative of θ. Because the entire target population can never be observed, there is sampling error and we are not certain of the summary estimate. Consequently, it is commonly accompanied by a confidence or credibility interval of values that are the most likely candidates for the true parameter value.

It is also possible to relax the assumption of a fixed effect and allow the estimates y_j to come from a parameter that has a distribution by itself. Looking at Figure 9.3 again, instead of one θ we now allow multiple θ_j and assume a normal distribution for it with mean μ and variance τ^2. The μ refers to a common component of the parameters θ_j for each study and τ^2 is the between-study variance component. This framework is referred to as a random effects model:

$$y_j \sim N\left(\theta_j, \sigma^2\right), \quad \theta_j \sim N\left(\mu, \tau^2\right). \tag{2}$$

If the fixed effect assumption is not appropriate and we assume random effects instead, calculating and providing a single summary estimate from a meta-analysis is philosophically speaking artificial because we already have acknowledged that there is no such single quantity that can represent all included studies. Even providing a confidence or credibility interval does not alter this situation as it only provides a range for a multitude of parameters, not a range for a single parameter. The random effects estimate is nevertheless commonly interpreted as an estimate that can represent all studies. Because the assumptions underlying the random effects model are less restrictive than those of the fixed effect model, the confidence intervals are wider and likely to be more representative of the uncertainty around the summary estimate.

9.4.3 Assessing the modeling assumptions

The underlying effect(s) assumption, i.e. fixed or random, has a big impact on the uncertainty of the inferences. Therefore, significant attention has been given to the development of methods to assess the appropriateness of the assumptions. These methods, graphical and numerical, generally assess whether the heterogeneity or inconsistency among the primary study estimates could have occurred from sampling error alone, in which case a fixed effect approach could be appropriate. If it is necessary to include a between-study variance component, a random effects model must be applied.

Before the heterogeneity is assessed in a statistical manner, one should first determine whether the clinical or empirical homogeneity between study characteristics, e.g. study populations, study settings, and design, is sufficient for a meta-analytical summary estimate to make sense from a clinical perspective. It is not impossible for a statistically homogeneous group of studies to be clinically too heterogeneous to be included in a meta-analysis. Although a summary estimate can almost always be calculated, it will not always be clinically meaningful.

Graphical assessments

The graphs that seem best suited to assess statistical heterogeneity appear to be the forest plot and the standardized residual histogram with a normal distribution overlay[24].

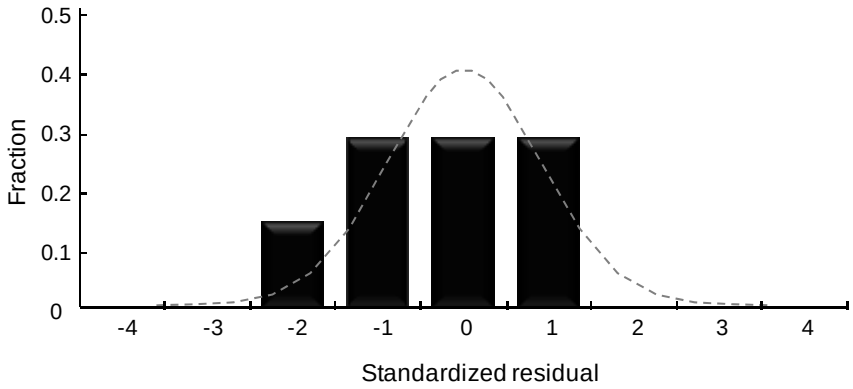


Figure 9.4 Standardized residual histogram

The standardized residual histogram (Figure 9.4), although straightforward to make and easy to interpret[24], is not commonly found in meta-analysis software[25]. The histogram plots the fractions of categorized standardized residuals (the individual estimate minus the summary estimate, divided by the standard error of

the individual estimate) in vertical bars. An overlay of a normal distribution can then be used to assess heterogeneity and departures from normality. Figure 9.4 indicates that there may not be much heterogeneity to worry about in the current data set. To produce the histogram in MIX, choose **'Exploration'** > **'View'** > **'Residual histogram'**.

The forest plot (Figure 9.5), first introduced in a meta-analysis context in 1982[26,27], displays the effect estimate of each included study with a square that is proportional to the study's weight; lines extending from the squares are proportional in length to the confidence intervals. The summary or meta-analysis estimate is often indicated by a diamond at the bottom. For exploratory purposes, MIX provides a (simple) forest plot without the diamond via **'Exploration'** > **'View'** > **'Simple forest plot'**. The forest plot in Figure 9.5 shows some evidence of heterogeneity, but the confidence intervals are generally wide and overlapping.

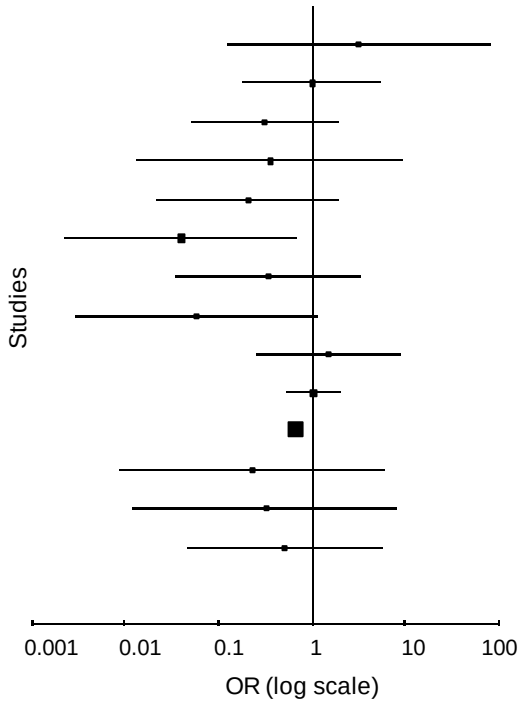


Figure 9.5 Forest plot

It is noteworthy that only 14 of the 22 studies are included in the graphs. Because the remaining eight found no events at all, they are excluded in a standard meta-analysis. Other commonly used graphs for assessing the extent of the between-study variation are the Galbraith plot and the L'Abbé plot. Both plots appear to be

interpreted with less consistency among raters than the forest plot and the aforementioned histogram[24]. The histogram, forest plot, Galbraith plot, and L'Abbé plot are all available in MIX. Appendix B.3 (graphical model exploration.r) contains code to reproduce these graphs in R.

Numerical assessments

Numerical assessments can involve statistics that fit in the fixed effect framework and those that fit in the random effects framework. Common statistics in the first category are the Q (mostly Cochran's Q[28,29]) and I^2 (I-squared)[30] statistics. Other tests, such as an (adjusted) Welch test[31] and an approximate ANOVA F test[32], have been proposed but are not integrated in most of the meta-analysis software[25].

The Q statistic is a weighted summary of residuals (individual estimates minus the summary estimate). It is chi-square distributed with k-1 degrees of freedom (k = number of studies), and it assesses whether all estimates could come from a single normal distribution. However, it has low power in meta-analyses with few studies[33,34] (giving insignificant results even when there is considerable heterogeneity). It also has a high type I error when many (large) studies are included[30] (i.e. falsely concluding that there is heterogeneity when there is none).

The I^2 quantifies inconsistency among study results and is a function of Q and its degrees of freedom:

$$I^2 = \left(Q - k - 1\right)/Q \,. \tag{3}$$

It is often multiplied by 100 and expressed in percentages. As such, it can be interpreted as the percentage of between-study variation that cannot be attributed to random sampling variation. The I^2 is a relative measure of between-study variability, i.e. relative to the within-study variation. For example, if two meta-analysis data sets have similar between-study variances but the sample sizes of the individual studies in one data set are larger (random sampling error smaller), the between-study variation will proportionally contribute more to the total variation in that data set and I^2 will be larger. It has therefore been reported that the I^2 should not be used as an absolute measure of heterogeneity[35].

The methods described above for assessing heterogeneity assume a fixed effect parameter θ for the calculations of Q. If there is no fixed effect, strictly speaking the Q (and consequently the I^2) does not make sense, since we are no longer interested in quantifying whether the studies could have come from a single

distribution. In a random effects framework, the between-study variance τ^2 (tau-squared) can be estimated in a number of ways. The most popular way, possibly due to its computational simplicity, is the version of DerSimonian and Laird[36]. However, other non-iterative versions such as those of Cochran[29] and DerSimonian and Kacker[37] have been used, as well as an iterative approach that has been shown to be statistically optimal[38].

The Q-statistic, I^2, and τ^2 are displayed in MIX via **'Exploration' > 'View' > 'Heterogeneity statistics'**. The statistics can be calculated in R as illustrated in Appendix B.4 (numerical model exploration.r). Heterogeneity exploration of the beta-blocker data set gives the following results: Q = 12.7 (P = 0.48), I^2 = 0% (95% CI: 0%-55%), and τ^2 = 0. This confirms what we already saw in the forest plot and the histogram, namely that much of the between-study variation may be attributed to sampling error. The τ^2 is zero, which means that the random effects model defaults back to the fixed effect model (without a variance component in the second part of Equation 2, θ_j has no distribution and becomes a single fixed value θ).

In the analysis section we will show that conclusions from the synthesis can be sensitive to the continuity correction, effect size, effect model, weighting method, and subgroups that are chosen. To prevent data dredging, it is therefore imperative that most of these analytical properties are pre-specified in the protocol.

9.5 Synthesis

9.5.1 Traditional aggregate-level synthesis

If the data appear suitable for synthesis, fixed effect or random effects models can be applied to calculate the summary estimate with a confidence interval. In a sense, this is the easiest part of a meta-analysis, because it requires little or no judgment; the judgments are made at the time of protocol writing and data exploration.

Summary estimates Y_Σ from traditional meta-syntheses of individual study estimates y_j, whether fixed or random, are weighted statistics:

$$Y_\Sigma = \frac{\sum_{j=1}^{k}\left(w_j \times y_j\right)}{\sum_{j=1}^{k} w_j}. \qquad (4)$$

Each study is weighted according to its statistical informativeness. There are multiple ways to express this informativeness and consequently there are a number of weighting methods. The most common method is the inverse-variance weighting:

$$w_j = \frac{1}{v_j} \ , \tag{5}$$

where v_j is the variance of the study. The standard error of the summary estimate is calculated by taking the inverse of the sum of the weights:

$$se_{\Sigma} = \sqrt{1 \Big/ \sum_{j=1}^{k} w_j} \ , \tag{6}$$

and confidence intervals are calculated in the usual fashion by using a standard deviate from a *z*-distribution or *t*-distribution.

Other fixed effect methods for binary outcome data are the Mantel-Haenszel[39] and Peto[40] approaches. Inverse variance weighting is most versatile as it can be applied to continuous outcomes as well as dichotomous outcomes. It is also useful if different outcome measures have been used by different studies and syntheses are performed on standardized measures such as standardized mean differences or correlation coefficients. The Peto method does not require continuity corrections and the Mantel-Haenszel method only for calculation of confidence intervals. The Peto and Mantel-Haenszel approaches may therefore be the methods of choice for data sets with sparse events[41]. If there are many imbalances in the sizes of the study arms, however, the standard Peto method can result in biased syntheses[42] and continuity corrections are required to solve that issue (see 9.6.2).

A number of syntheses can be performed in MIX via **'Synthesis'** > **'View'** > **'Synthesis'**. The association measure, synthesis model, and weighting method can be changed quickly via the **Synthesis'** > **'Change'** > **…** menus. The fixed effect result of the Mantel-Haenszel odds ratio synthesis is 0.63 (95% CI: 0.52, 0.76). The DerSimonian-Laird estimate of τ^2 is 0 and the random effects synthesis is identical to the inverse-variance fixed effect synthesis: 0.65 (95% CI: 0.53, 0.79). Thus, these syntheses suggest that beta-blocker therapy is associated with a 35% to 37% decreased risk of non-fatal myocardial infarction[21].

The results of a synthesis are often illustrated with an annotated forest plot. The plot not only shows how the individual effect sizes evolve into a summary effect size, but commonly also gives the study identifiers, some raw data, study weights, and details of heterogeneity statistics and the summary estimate. The top part of Table 9.2 gives an overview of results from different weighting methods in traditional syntheses. The R code for the syntheses is given in Appendix B.5 (traditional synthesis.r).

Table 9.2 Comparison of results from various methods

Method	*k*	OR (95% CI)
Traditional synthesis		
Inverse-variance	14	0.65 (0.53, 0.79)
Mantel-Haenszel	14	0.63 (0.52, 0.76)
Peto	14	0.63 (0.52, 0.76)
DerSimonian-Laird	14	0.65 (0.53, 0.79)
Regression synthesis		
Linear fixed effect	14	0.65 (0.53, 0.79)
Linear random effects	14	0.65 (0.53, 0.79)
Logistic fixed contingency table	22	0.60 (0.50, 0.73)
Logistic hierarchical individual	22	0.62 (0.51, 0.76)
Bayesian synthesis		
Closed form	14	0.65 (0.53, 0.79)
Binomial fixed	22	0.62 (0.51, 0.77)

Abbreviations: *k* = number of included studies, OR = odds ratio.

9.5.2 Linear regression synthesis

The simplest regression method for synthesis of multiple study data is a fixed effect weighted linear regression of the effect size (if applicable with an appropriate link function) on an intercept β_0 only:

$$y_j = \beta_0 + \varepsilon_j . \tag{7}$$

The error terms ε_j are assumed to be independent and identically distributed, with a normal distribution, and the weights are defined as the inverse of the variance of each study. This can be expressed as follows:

$$y_j \sim N(\beta_0, \sigma^2) , \tag{8}$$

which is just a different way of writing Equation 1. It shows that the fixed effect regression model is in fact the same as the classical inverse variance synthesis method, with the intercept estimating the parameter θ. The advantage of

formulating the synthesis in a regression model is that later in the analysis phase, additional covariates can be added to the model to explain between-study variation in the effect sizes. Software that can do ordinary least squares regression can fit this regression model. However, the standard error for the intercept is biased for meta-analytical purposes[43] and needs to be adjusted to get the correct confidence intervals for the synthesis's measures of association:

$$se_{adj}(\beta_0) = se(\beta_0) / RMS \,, \tag{9}$$

where *RMS* is the residual mean square from the ANOVA table.

MIX can fit the above model without covariates or with a single covariate via **'Analysis' > 'View' > 'Meta-regression'**. The results are identical to those of the inverse variance synthesis described earlier (Table 9.2). The same model can be fitted in R (Appendix B.6: fixed effect linear regression.r).

The intercept-only fixed effect regression model can be expanded by including a random effect. It is beyond the scope of this paper to discuss the various approaches to random effects regression analyses and we will focus on the complement of the classical random effects meta-analysis in the regression framework. Contrary to the fixed effect regression, where the total between-study variation is assumed accounted for by sampling and covariates, the random intercept regression model takes residual variability into account by including a random effect u_j:

$$y_j = \beta_0 + u_j + \varepsilon_j \,. \tag{10}$$

The u_j is then assumed to be normally distributed with mean 0 and unknown variance τ^2:

$$u_j \sim N(0, \tau^2) \,. \tag{11}$$

This is the regression analogy to the random effects model in Equation 2. The random effects model cannot be fitted in MIX but can be analyzed in R as illustrated in Appendix B.7 (random effects linear regression.r). The results may depend on what estimator for τ^2 is used. For the current data set it doesn't matter since all estimates for τ^2 are 0 and the odds ratio summary estimate is identical to the standard inverse-variance synthesis results: 0.65 (95% CI: 0.53, 0.79).

9.5.3 Logistic regression synthesis

A disadvantage of the linear regression models described above is that continuity corrections are required to fit the models on (effect sizes from) sparse event data. Direct modeling of the events with logistic regression eliminates the need for such corrections. Although this is most intuitive in a Bayesian framework, as demonstrated in the next section, there are also ways to apply a logistic model to meta-analytical data with standard statistical software.

A fixed effect weighted logistic model can be fitted to semi-summarized data[44]. The data must be transformed to hold *k* units of analysis per included study, where *k* is the number of cells in the contingency table of the particular study. For example, in a randomized trial with one index treatment and one reference treatment, the contingency table would be a normal two-by-two table with four cells. The subsequent four additional rows per study in the aggregate meta-analysis data set will hold data for two additional columns: a column for the binary outcome (coded 0 or 1) and a column for the binary arm assignment (coded 0 or 1). One additional column holds the number of subjects in each study that correspond to the outcome and assignment characteristics of each row. The values in the latter column will be used as weights in the logistic regression. Table 9.3 shows the results of this data transformation for the first two studies in the beta-blocker data set.

Table 9.3 Data set transformation for contingency table logistic regression

Id	g1.r1	g1.r0	g1.n	g0.r1	g0.r0	g0.n
Jacobsen I	1	17	18	0	18	18
Lai	3	27	30	3	27	30

Id	Assignment (g)	Outcome (r)	Subjects
Jacobsen I	1	1	1
Jacobsen I	1	0	17
Jacobsen I	0	1	0
Jacobsen I	0	0	18
Lai	1	1	3
Lai	1	0	27
Lai	0	1	3
Lai	0	0	27

MIX has facilities to transform the data set via: **'Main' > 'Data wizard' > 'Transform data' > 'Aggregate to contingency'**, but it cannot perform the weighted logistic regression itself. The new data set created by MIX in Excel can, however, be exported to CSV file format and loaded in R. The analysis in R can

be performed by the *glm* package, as illustrated in Appendix B.8 (logistic contingency table regression.r). For the beta-blocker data set the results are now slightly different with an odds ratio of 0.60 (95% CI: 0.50, 0.73). This is because all 22 studies can be modeled in the binomial-logistic model. In the traditional syntheses and linear models, studies with no events at all were excluded and only 14 studies remained available for analysis (Table 9.2).

Another approach to logistic meta-regression is based on a hierarchical regression model that incorporates the nested structure of the data (individuals nested in studies). First, the data set needs to be transformed to hold the individual outcomes for each individual subject and requires an additional column to indicate in which study each subject was enrolled. MIX has facilities to transform the aggregate-level data set to an individual-level data set: **'Main' > 'Data wizard' > 'Transform data' > 'Aggregate to individual'**. The resulting per-subject level data set can be exported as a CSV file and analyzed in R with multilevel logistic regression. This is essentially an individual patient data analysis, except that no individual covariates are included. The results lie between those of the traditional syntheses and the logistic contingency table regression with an odds ratio of 0.62 (0.51, 0.76). Appendix B.9 (logistic individual regression.r) illustrates this kind of analysis in R.

9.5.4 Bayesian synthesis

There are two major views of probability; one is referred to as Frequentist and the other as Bayesian. The methods presented so far fit in the Frequentist framework. In the Frequentist framework, scientific probability is formally defined as the frequency with which the encountered study results would occur in a hypothetical sample of such studies. The ranges of the Frequentist confidence intervals therefore do not directly infer to the underlying parameters, but rather indicate a possible parameter range that is compatible with the study's results.

Another approach is to define probability not as a trait of data, but as a degree of belief, potentially different for everyone. Experience is then seen as a way of updating these ideas of reality: each time we experience an event we revise our belief about what caused this event in light of that particular experience. Formalization of this process in science requires usage of some probability mathematics called 'Bayes' rule':

$$P(\theta \mid Data) \propto P(Data \mid \theta) P(Data). \qquad (12)$$

The first probability in Equation 11 is often called the posterior probability and it is proportional to the likelihood (the second probability in the equation) multiplied by the prior probability. The prior probability is the probability that we assign to parameter values before we integrate knowledge we have gained from a new study. The likelihood is the probability of the study results conditional on parameter assumptions. Because Bayesian analyses estimate underlying parameters directly, there is no need for *P* values and the interpretation of confidence intervals is much more straightforward. The latter refer directly to parameters. Although the Bayesian framework is conceptionally very simple (updating prior knowledge to posterior knowledge through integration of the likelihood), the analytical part can be daunting and most comprehensive analyses require Markov Chain Monte Carlo (MCMC) simulation.

MIX can do a very rudimentary Bayesian analysis on effect sizes, based on the assumption that the effect sizes are normally distributed (**'Synthesis' > 'View' > 'Bayesian synthesis')**. For the beta-blocker data, this can be done with the logarithm of the odds ratio as effect size. With a standard continuity correction, the same 14 studies as included in the traditional Frequentist syntheses are used. With a non-informative prior, the results are identical to these syntheses (Table 9.2). Nevertheless, this rudimentary Bayesian approach in MIX does not take advantage of the fact that continuity corrections can be avoided by modeling the event rates in the treatment group ($g1.r1_j$) and control group ($g0.r1_j$) directly:

$$\begin{aligned} g1.r1_j &\sim Bin\left(g1.p_j, g1.n_j\right), \\ g0.r1_j &\sim Bin\left(g0.p_j, g1.n_j\right). \end{aligned} \tag{13}$$

The binomial probabilities in the control group ($g0.p_j$) and treatment group ($g1.p_j$) can then be defined linearly via a logit transform:

$$\begin{aligned} \log\left(\frac{g0.p_j}{1-g0.p}\right) &= bm_j, \\ \log\left(\frac{g1.p_j}{1-g1.p}\right) &= bm_j + tm. \end{aligned} \tag{14}$$

The bm_j is the baseline model (log odds for the control group) and *tm* refers to the treatment model (log odds ratio). To make our example concise and not overly complicated when it comes to specifying prior distributions, Equation 13 presents

a fixed effects approach where the treatment model is constant for all studies (*tm* has no subscript *j*). The priors are specified as non-informative (very wide) normal distributions:

$$bm_j \sim N(0,10000),$$
$$tm \sim N(0,10000). \quad (15)$$

Random effects approaches would model tm_j or bm_j (or both) with hyper-parameters, that are then given distributions. Interested readers are referred to two well-written introductory chapters on Bayesian evidence synthesis and analysis in books by Spiegelhalter et al.[45] and Sutton et al[3].

The OpenBUGS software was used to perform this Bayesian meta-synthesis (Appendix B.10: Bayesian hierarchical binomial.odc). With the uninformative priors on the parameters, the results are almost identical to the results from the logistic regression of the individual-level data set (Table 9.2).

9.6 Evaluation

9.6.1 Basic synthesis characteristics

We have stated previously that it is important to specify the major characteristics of the synthesis, e.g. the type of effect size, the effect model, and the weighting method, in the protocol of the systematic review. Although the synthesis should be performed accordingly, it should be evaluated how sensitive the results are to alternative approaches. For example, an odds ratio may be reported in the main results section, but the impact of changing from odds ratio to risk ratio or risk difference may be provided in a table with other sensitivity analyses. Basic evaluations can be performed in MIX via: **'Evaluation' > 'Change'** and **'Synthesis' > 'Change'**.

9.6.2 Continuity corrections

So far, the data have been analyzed with the traditional constant correction for continuity of adding 0.5 to the event values of a study with zero events in one of the study arms. Although convenient and easy to implement in software, the constant corrections can cause bias if there are imbalances in study arms[42,46]. For example, in the beta-blocker data set four studies (Zaug, Stome, Raby, and Miller) have substantial imbalances. The occurrence of the bias is best seen in a study with no events in both arms, like the Miller study. Adding 0.5 to the event values of the treatment arm with 368 subjects and adding the same value to the other arm

with 180 subjects introduces an association (odds ratio of 0.49) that is neither neutral nor apparent from the data. The same is true for smaller corrections like 0.01.

Variable continuity corrections have been proposed that take the number of subjects in the study arms into account[46]. The corrections add a value proportional to the reciprocal of the size of the opposite treatment group. With R being the ratio of group sizes and S being the normalization factor to which the corrections for event or no event cells in a single study arm sum up (commonly set to 1), the variable correction adds a factor of R/(S*(R+1)) to the larger group and 1/(S*(R+1)) to the other group. The study arm correction is equal to the constant correction if studies have equal size study arms. Applying this correction to the Miller study adds 0.67 to the event values in the study arm with 368 subjects and 0.33 to the event values in the other arm. The resulting odds ratio is 1.

Studies without any events are commonly excluded from meta-analyses because they provide no information about the magnitude of the effect sizes[41,46]. On the other hand, these studies do provide relevant information by showing that event rates for both the intervention and control groups are low[47,48]. To integrate this information, it is possible to apply continuity corrections to studies without events in either study arm.

In MIX, all of the above-described continuity corrections can be explored via **'Edit' > 'Numerics' > 'Binary data' > 'Continuity corrections'**. Table 9.4 gives an overview of the impact of different continuity corrections on the results of the inverse variance synthesis of the studies in the beta-blocker data set. Although the impact is minor in this analysis, it can be substantial[42] and methods should be pre-specified in the protocol to prevent fishing for significance.

Table 9.4 Impact of continuity corrections

Method	*k*	OR (95% CI)
Studies without events excluded		
Constant correction (0.5)	14	0.65 (0.53, 0.79)
Constant correction (0.01)	14	0.66 (0.54, 0.81)
Variable correction	14	0.65 (0.53, 0.79)
Correction applied to studies without events		
Constant correction (0.5)	22	0.65 (0.54, 0.79)
Constant correction (0.01)	22	0.66 (0.54, 0.81)
Variable correction	22	0.65 (0.54, 0.79)

Abbreviations: *k* = number of studies; OR = odds ratio; IV = inverse variance; MH = Mantel-Haenszel.

9.6.3 Subgroups

If heterogeneity is substantial, it may be indicated to create subgroups of studies with similar characteristics. Although one could explore the numerical clustering of studies in the data set and decide on subgroups during the analysis, it has been shown that sub-group analyses can show spurious effects that are based on chance and not on true between-study variation[49,50]. It is therefore imperative that sub-groups are pre-specified in the protocol based on clinical reasoning[51]. The authors of the beta-blocker meta-analysis made subgroups based on outcomes (e.g all-cause mortality, cardiovascular mortality, non-fatal myocardial infarction, non-fatal stroke) and the risk of bias (high and low risk). For non-fatal myocardial infarction as outcome, stratification of studies based on risk of bias shows a substantial data trend. Synthesis of studies with high risk of bias gives an odds ratio of 0.34 (0.15, 0.78), while synthesis of studies with low risk of bias results in an odds ratio that is much less extreme: 0.67 (0.55, 0.83).

9.6.4 Extending regression models with covariates

Reasons for the between-study variation can be analyzed by extending the regression techniques discussed in the synthesis section above. For example, the linear regression model of Equation 6 can be extended with covariates as follows:

$$y_j = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \ldots + \beta_m x_{jm} + u_j + \varepsilon_j \,. \qquad (16)$$

This type of regression shows the association between covariates and the effect size as well as the impact of adding covariates on the overall effect. However, the units of analysis in these regressions are the studies and there is a high chance of finding spurious associations if many covariates are tested. Like subgroup analyses, it is therefore advisable to limit the number of covariates and pre-specify them in the protocol of the systematic review[52]. An alternative to using multiple predictors is to use baseline risk as aggregate representative of these factors and include it as single covariate in the model. If this approach is taken, it is nevertheless advisable to use a random effects model and include a parameter representing residual heterogeneity[53]. Appendix B.11 (linear random effects covariate.r) gives an example with a fictional continuous covariate in R.

9.6.5 Incorporating study quality

Many systematic reviews involve some kind of assessment of quality (Delphi) or risk of bias (Cochrane) for each included study. It is therefore attractive to also include such an assessment as a component of influence in the synthesis. If the

risk-of-bias assessment results are dichotomized, one can perform a simple subgroup analysis as presented in section 9.6.3. If the risk-of-bias assessment is used to produce continuous scores for each study, a number of approaches can be taken. The simplest method is to create bias-adjusted weights baw_j by multiplying the original inverse variance weights w_j with a risk-of-bias score rbs_j[54,55]. With these weights, a standard weighted synthesis can be performed:

$$Y_{\Sigma} = \frac{\sum_{j=1}^{k} \left(baw_j \times y_j \right)}{\sum_{j=1}^{k} baw_j}. \tag{17}$$

The standard error of the summary estimate is not calculated by taking the inverse of the sum of the weights as in Equation 6, but rather in the following manner:

$$se_{\Sigma} = \frac{\sum_{j=1}^{k} \left(baw_j{}^2 \times w_j \right)}{\left(\sum_{j=1}^{k} baw_j \right)^2}. \tag{18}$$

However, this affects the width of the confidence interval of the synthesis's result in a manner that lacks statistical justification[54,56].

An alternative is proposed by Doi and Thalib[57]. They first create a proportional risk-of-bias score $prbs_j$ by dividing each score by the maximum attainable score. Next, a weight distributor function d is introduced:

$$d_j = \frac{1}{w_j} \times \left(\sum_{j=1}^{k} \left(\frac{w_j - \left(w_j \times prbs_j \right)}{k-1} \right) - \frac{w_j - \left(w_j \times prbs_j \right)}{k-1} \right). \tag{19}$$

This function is used to create proportional bias-adjusted weights $wpba_j$ as follows:

$$wpba_j = w_j \times \left(d_j + prbs_j \right), \tag{20}$$

and these weights are used to calculate a risk-of-bias-adjusted estimate as in Equation 17.

An alternative method is to include the risk-of-bias score as a covariate in a regression analysis as described in 9.6.4. In MIX, risk-of-bias scores can be integrated in the analysis with the Doi and Thalib method or with univariable regression via **'Analysis' > 'Change' > 'Integrate risk-of-bias assessment'**. Univariable or multivariable regression can be performed in R as shown in Appendix B.11 (linear random effects covariate.r).

9.6.6 Dissemination bias

Evidence dissemination is the process by which scientific data are transformed into evidence for the public domain. At various stages of this process, selectivities can occur that over or under-disseminate certain evidence and potentially result in dissemination bias. A variety of methods are applied in this context. We make a distinction between methods that attempt to quantify data trends due to selectivity and those that attempt to quantify the bias resulting from this selectivity. The data trends that are commonly attributed to selective dissemination of evidence are so-called small-study effects[58,59]. The term refers to the association that occurs between the effect size and the precision because small, negative, and insignificant studies are suppressed in the data set. It is noteworthy, however, that such data trends can also be caused by genuine heterogeneity, chance, or issues related to study quality[60].

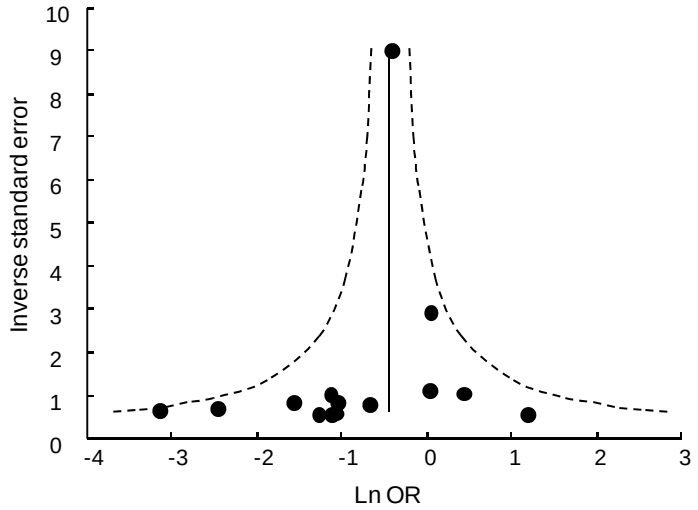


Figure 9.5 The funnel plot

Graphical assessments

The most common graphical assessment of data trends that may be due to dissemination selectivity is the funnel plot, introduced by Light and Pillemer in 1985[61]. It typically has a measure for effect size on the x-axis and a measure related to the within-study variance (e.g. the standard error, precision, or the sample size) on the y-axis. Each study is represented by a single equally-sized dot. Under most circumstances, there are relatively more small studies (with larger variance) than big studies in a meta-analytical data set and the smaller studies have estimates that are more scattered and further removed from the summary estimate. This creates a funnel-like symmetrical distribution of the dots in the plot. If small studies are relatively 'missing' on one side of the plot, a subsequent asymmetry in the low precision area can occur. Figure 9.5 shows the funnel plot for the beta-blocker data set. There is some evidence of small-study effects with some missingness in the bottom right area. In MIX, a number of funnel plots can be made via **'Analysis' > 'View' > 'Funnel plots'**. Examples of how to make funnel plots in R are provided in Appendix B.12 (funnel plots.r).

Numerical tests for data trends due to selectivity

A number of tests exist to assess the data trends numerically. The rank correlation test by Begg[62] and Egger's regression test[60] are available in most software packages and we will limit our description to these two methods.

The rank correlation test[62] examines the association between standardized effect sizes and their (within-study) variances. Smaller studies with higher variances will generally have larger empirical values and the test assumes that publication bias will induce a correlation between these two factors (e.g. that there are more positive studies with a larger variance). The correlation is tested by means of Kendall's τ, which can be normalized and tested for significance by means of its z value. If the P value is below the accepted alpha level, the null hypothesis of no correlation (absence of publication bias) is rejected. The method requires no modeling assumptions but suffers from lack of power and hence a liberal significance level (alpha) is suggested[59].

A regression test developed by Egger et al.[60] to assess small-study effects is based on an ordinary least squares regression of each study's standardized effect size z_j (the effect size divided by its standard error) on the precision x_j (inverse of the standard error).

$$z_j = \beta_0 + \beta_1 x_{1j} . \qquad (21)$$

The intercept of the regression line is used to measure asymmetry around the standardized pooled value 0. Deviance from 0 is regarded as a sign of publication bias with *P* values and confidence intervals obtained as usual in least squares regressions.

In MIX, the rank correlation and regression tests can be performed via the **'Analysis' > 'View' > 'Dissemination bias'** menu. The code to do the tests in R is provided in Appendix B.13 (bias tests.r). For the beta-blocker data set, the *z* value from Kendall's τ is 0.77 with a *P* value of 0.44 and the test for 0 intercept with Egger's method yields a *P* value of 0.22. This indicates that the small-study effects are small.

Quantifying potential bias

There are a number of approaches that quantify the potential dissemination bias in the meta-analytical results. Most promising are selection models[63] that model the selectivity processes in the evidence dissemination explicitly and use this modeling to produce summary estimates that are corrected for the selectivity. Although programs to run some of the models in R are available[64], they require iterative methods that do not always converge when common effect sizes like the odds ratio or risk ratio are used. Because of their complexity, we will not further illustrate the methods here.

The trim-and-fill method[65,66] is a simple method for estimating and imputing studies that may be missing due to selective dissemination. This is followed by a meta-synthesis that includes the imputed studies. A funnel plot can be produced that adds the imputed studies (usually dots that are not filled), as well as an additional vertical line that indicates the summary effect when the imputed studies are included in the synthesis. When the method is applied to the beta-blocker data set in a fixed effect inverse variance analysis, four studies are imputed. This may seem substantial, but using the number of imputed studies to judge whether there are small-study effects can be misleading. The imputation of the studies shifts the effect size from 0.65 (95% CI: 0.53, 0.79) to 0.67 (95% CI: 0.55, 0.81). This shift is minimal and confirms the results from the rank correlation and regression tests.

Figure 9.6 shows the imputed studies and the shift of the summary estimate in a trim-and-fill funnel plot. The trim-and-fill assessment can be performed in MIX via **'Analysis' > 'View' > 'Dissemination bias'** and the plot can be produced via **'Analysis' > 'View' > 'Trim-and-fill plot'**. Appendix B.14 (trim and fill.r) contains code to reproduce the results in R.

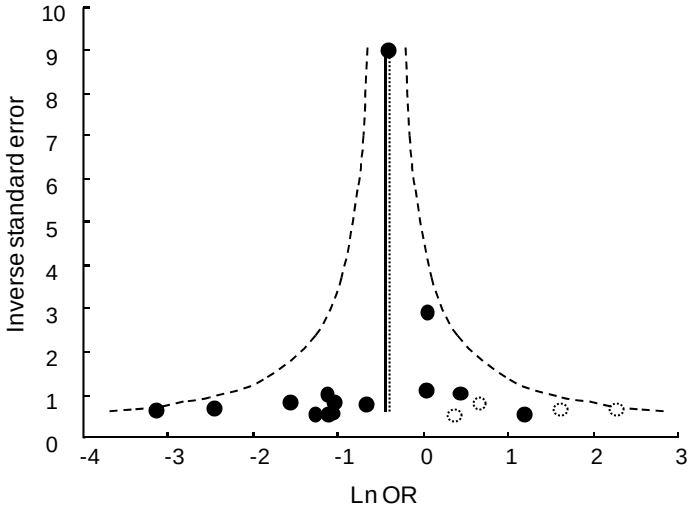


Figure 9.6 The trim-and-fill plot

9.7 Software used in this paper

The meta-analytical output for this paper was produced with MIX 2.0, R 2.8.1, and OpenBUGS 3.0. MIX is an add-in for Excel 2007 and can be downloaded from *http://www.mix-for-meta-analysis.info*. It creates a graphical user interface inside Excel and is capable of almost all analyses presented in this paper. R is a language and Open Source environment for statistical computing and graphics and is available from *http://www.r-project.org*. There is very little that cannot be done in R, but performing analyses involves writing code. OpenBUGS is Open Source software for Bayesian analysis of complex statistical models using Markov Chain Monte Carlo (MCMC) methods and can be downloaded at *http://mathstat.helsinki.fi/openbugs/*. Like R it is very flexible, but it requires some knowledge of statistical programming.

9.8 Conclusion

We have presented a comprehensive overview of methods in meta-analysis and structured them in a framework of exploration, synthesis, and evaluation. The overview is by no means complete and some recent developments in the meta-analysis of multiple treatments and individual patient data have been omitted in favor of practical applicability. The manuscript nevertheless provides details about a wide range of topics such as fixed and random effects assumptions, traditional methods for meta-synthesis, regression models, Bayesian approaches, continuity corrections, and assessment of the potential risk of dissemination bias and its impact on synthesis results. Distinguishing exploration, synthesis, and evaluation

in a meta-analytical project should encourage meta-analysts to go beyond the synthesis; exploration of the data set and evaluation of reasons for between-study variation and the sensitivity of synthesis results to alternative analytical approaches should be part of the standard meta-analysis repertoire.

Acknowledgments

We would like to thank Harukazu Tsuruta for his comments on the manuscript and his help with the statistical sections.

Conflict of interest

The first author is the primary programmer of the MIX software. MIX is offered as a free download for academic purposes. All registration payments from private and commercial users are used to maintain and develop the program.

References

1. Egger M, Davey Smith G, Altman D. *Systematic Reviews in Health Care: Meta-analysis in Context.* London: BMJ Publishing Group; 2001.
2. Glasziou P, Irwig L, Bain C, Colditz G. *Systematic Reviews in Health Care: A Practical Guide*. Cambridge: Cambridge University Press; 2001.
3. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. *Methods for Meta-Analysis in Medical Research*. Chichester: Wiley; 2000.
4. Sutton AJ, Higgins JP. Recent developments in meta-analysis. *Stat Med.* Feb 28 2008;27(5):625-650.
5. Khoshdel A, Attia J, Carney SL. Basic concepts in meta-analysis: A primer for clinicians. *Int J Clin Pract.* Oct 2006;60(10):1287-1294.
6. Poincare JH. *Science and Hypothesis*. London and Newcastle-on-Cyne: The Walter Scott publishing Co.; 1902.
7. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat Med.* Feb 28 2002;21(4):589-624.
8. Stevens RD, Wu CL. Strengths and limitations of meta-analysis. *J Cardiothorac Vasc Anesth.* Feb 2007;21(1):1-2.
9. Lyman GH, Kuderer NM. The strengths and limitations of meta-analyses based on aggregate data. *BMC Med Res Methodol.* 2005;5(1):14.
10. DeMaria AN. Meta-analysis. *J Am Coll Cardiol.* Jul 15 2008;52(3):237-238.
11. Wachter K. Disturbed by Meta-Analysis? *Science.* 1988;241(4872):1407-1408.
12. Egger M, Smith GD, Sterne JA. Uses and abuses of meta-analysis. *Clin Med.* Nov-Dec 2001;1(6):478-484.

13. Rothstein HR. Publication bias as a threat to the validity of meta-analytic results. *J Exp Criminol.* 2008;4(1):61-81.
14. Eysenk HJ. An exercise in mega-silliness. *Am Psychol.* 1978;33:517.
15. Feinstein AR. Meta-analysis: statistical alchemy for the 21st century. *J Clin Epidemiol.* 1995;48(1):71-79.
16. van Houwelingen HC. The future of biostatistics: expecting the unexpected. *Stat Med.* Dec 30 1997;16(24):2773-2784.
17. Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. *Int J Epidemiol.* Mar 1989;18(1):269-274.
18. Morgenstern H. Uses of ecologic analysis in epidemiologic research. *Am J Public Health.* Dec 1982;72(12):1336-1344.
19. Jones AP, Riley RD, Williamson PR, Whitehead A. Meta-analysis of individual patient data versus aggregate data from longitudinal clinical trials. *Clin Trials.* Feb 2009;6(1):16-27.
20. Sutton AJ, Kendrick D, Coupland CA. Meta-analysis of individual- and aggregate-level data. *Stat Med.* Feb 28 2008;27(5):651-669.
21. Bangalore S, Wetterslev J, Pranesh S, Sawhney S, Gluud C, Messerli FH. Perioperative beta blockers in patients having non-cardiac surgery: a meta-analysis. *Lancet.* Dec 6 2008;372(9654):1962-1976.
22. Altman DG. *Practical Statistics for Medical Research*. New York: Chapman & Hall / CRC; 1990.
23. D'Agostini G. *Bayesian Reasoning in Data Analysis*. Singapore: World Scientific Publishing; 2003.
24. Bax L, Ikeda N, Fukui N, Yaju Y, Tsuruta H, Moons KG. More than numbers: the power of graphs in meta-analysis. *Am J Epidemiol.* Jan 15 2009;169(2):249-255.
25. Bax L, Yu LM, Ikeda N, Moons KG. A systematic comparison of software dedicated to meta-analysis of causal studies. *BMC Med Res Methodol.* 2007;7(1):40.
26. Lewis JA, Ellis SH. A statistical appraisal of postinfarction betablocker trials. *Prim Cardiol.* 1982(supp1):31-37.
27. Lewis S, Clarke M. Forest plots: trying to see the wood and the trees. *BMJ.* Jun 16 2001;322(7300):1479-1480.
28. Cochran WG. Problems arising in the analysis of a series of similar experiments. *J Roy Stat Soc.* 1937;4(Supplement):102-118.
29. Cochran WG. The combination of estimates from different experiments. *Biometrics.* 1954; 10:101-129.
30. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ.* Sep 6 2003;327(7414):557-560.
31. Welch BL. On the comparison of several mean values: An alternative approach. *Biometrika.* 1951;38(330-336).
32. Asiribo O, Gurland J. Coping with variance heterogeneity. *Commun Stat Theory Method.* 1990;19:4029-4048.

33. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. *Stat Med.* Apr 30 1998;17(8):841-856.
34. Paul SR, Donner A. Small sample performance of tests of homogeneity of odds ratios in K 2 x 2 tables. *Stat Med.* Jan 30 1992;11(2):159-165.
35. Rucker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I(2) in assessing heterogeneity may mislead. *BMC Med Res Methodol.* 2008;8:79.
36. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controll Clin Trials.* 1986;7:177-188.
37. DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. *Contemp Clin Trials.* Feb 2007;28(2):105-114.
38. Paule RC, Mandel J. Consensus values and weighting factors. *J Res Natl Bur Stand.* 1982;87:377-385.
39. Robins J, Greenland S, Breslow NE. A general estimator for the variance of the Mantel-Haenszel odds ratio. *Am J Epidemiol.* 1986;124:719-723.
40. Yusuf S, Peto R, Lewis J, Colins R, Sleight P. Beta blockade during and after myocardial infarction. An overview of randomized trials. *Progress Cardiovasc Dis.* 1985;27:335-371.
41. Bradburn MJ, Deeks JJ, Berlin JA, Russell Localio A. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Stat Med.* Jan 15 2007;26(1):53-77.
42. Diamond GA, Bax L, Kaul S. Uncertain effects of rosiglitazone on the risk for myocardial infarction and cardiovascular death. *Ann Intern Med.* Oct 16 2007;147(8):578-581.
43. Hedges LV. Fixed effect models. *The Handbook of Research Synthesis*. New York: Russel Sage Foundation; 1994.
44. Agency for Healthcare Research and Quality. Meta-regression Approaches: What, Why, When, and How? http://www.ahrq.gov/clinic/tp/metaregtp.htm. Accessed April, 2009.
45. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Chichester: Wiley; 2004.
46. 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.* May 15 2004;23(9):1351-1375.
47. Friedrich JO, Adhikari NK, Beyene J. Inclusion of zero total event trials in meta-analyses maintains analytic consistency and incorporates all available data. *BMC Med Res Methodol.* 2007;7:5.
48. Sankey SS, Weissfeld LA, Fine MJ, Kapoor W. An assessment of the use of the continuity correction for sparse data in meta-analysis. *Commun Stat Simul Comput.* 1996;25:1031-1056.
49. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA.* Jul 3 1991;266(1):93-98.

50. Counsell CE, Clarke MJ, Slattery J, Sandercock PA. The miracle of DICE therapy for acute stroke: fact or fictional product of subgroup analysis? *BMJ.* Dec 24-31 1994;309(6970):1677-1681.
51. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med.* Jan 1 1992;116(1):78-84.
52. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? *Stat Med.* Jun 15 2002;21(11):1559-1573.
53. Sharp SJ, Thompson SG. Analysing the relationship between treatment effect and underlying risk in meta-analysis: comparison and development of approaches. *Stat Med.* Dec 15 2000;19(23):3251-3274.
54. Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. *J Clin Epidemiol.* Mar 1992;45(3):255-265.
55. Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet.* Aug 22 1998;352(9128):609-613.
56. Tritchler D. Modelling study quality in meta-analysis. *Stat Med.* Aug 30 1999;18(16):2135-2145.
57. Doi SA, Thalib L. A quality-effects model for meta-analysis. *Epidemiology.* Jan 2008;19(1):94-100.
58. Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Stat Med.* Oct 30 2006;25(20):3443-3457.
59. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *J Clin Epidemiol.* Nov 2000;53(11):1119-1129.
60. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ.* Sep 13 1997;315(7109):629-634.
61. Light RJ, Pillemer DB. *Summing up*. Cambridge: Harvard University Press; 1985.
62. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics.* Dec 1994;50(4):1088-1101.
63. Copas JB, Shi JQ. A sensitivity analysis for publication bias in systematic reviews. *Stat Methods Med Res.* Aug 2001;10(4):251-265.
64. Carpenter JR, Schwarzer G. The copas package. Statistical methods to model and adjust for bias in meta-analysis. *The R Project for Statistical Computing.* 2009.
65. Duval SJ, Tweedie RL. A non-parametric 'trim and fill' method of accounting for publication bias in meta-analysis. *J Am Stat Assoc.* 2000;95:89-98.
66. Duval SJ, Tweedie RL. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics.* Jun 2000;56(2):455-463.

10

SUMMARY

10.1 Introduction

In **Chapter 1**, we introduce the term meta-analysis. We show how it should be part of a systematic review and how it differs from vote-counting or qualitative approaches. Medical research can be categorized in causal and descriptive research and we describe how this dissertation focuses on the methods for meta-analysis of data from causal medical research. Meta-analyses are not always feasible and many have criticized the indiscriminate application of this tool. We argue that, if used properly, meta-analysis is indispensable if one is to quantitatively integrate prior evidence into clinical decision-making and clinical practice guidelines.

10.2 Methodology

The methodology part reports a number of methodological projects related to causal meta-analysis. In **Chapter 2**, we explore how experienced or inexperienced authors of meta-analyses are. We retrieved all citations in Medline (PubMed) that were classified as a meta-analysis in 2008. Using the citation information, we performed a systematic exploration of other Medline-registered meta-analyses of the first and the second or last authors. We found that roughly 28% of the first authors had been involved in a meta-analysis as primary investigator prior to their meta-analysis in 2008, while approximately 41% had previously been involved in a meta-analysis as co-author. The second and last authors combined were commonly more experienced, 32% had never participated in a meta-analysis and 59% had not been involved in one as a first author. We concluded that primary investigators of meta-analyses are often not experts in meta-analysis and that co-authors are generally more experienced. This has implications for how meta-analyses are (re)viewed and how educational materials and software for meta-analysts are designed.

Chapter 3 is an essay about selectivities and biases in the dissemination of evidence. The term publication bias is widely used to describe the tendency of investigators, reviewers, and editors to submit or accept manuscripts for publication based on the direction or strength of the study findings. It is not commonly recognized that the term is inaccurate. Evidence dissemination can be viewed as a multi-staged process with dissemination alternatives at distinct points in time. Selectivities in reporting (of outcomes as well as of entire studies), publication, and inclusion of evidence can lead to a number of biases. We propose to use the terms reporting bias, publication bias, and inclusion bias, respectively. The aggregate of these biases is best described by the term dissemination bias. The common usage of the term publication bias to represent all biases thus constitutes

a ‘pars pro toto’ – a part to describe the whole - and is therefore inappropriate in our view. We also distinguish meta-analytical methods that make a judgment on whether dissemination selectivity has likely occurred or not, and methods that attempt to quantify or correct for the bias that may have been induced by the dissemination selectivity. We conclude that a careful use of terms can facilitate the understanding of problems that occur in the dissemination of evidence.

Graph assessments are widely used to identify or rule out heterogeneity and dissemination selectivity. A variety of graphs are available for this purpose. To date, however, there has been no comparative evaluation of the performance of these graphs. In **Chapter 4**, we assess the reproducibility and validity of the graph ratings. We simulated 100 meta-analyses from four scenarios that covered situations with and without heterogeneity and selectivity. From each meta-analysis we produced 11 graphs (box plot, weighted box plot, standardized residual histogram, normal quantile plot, forest plot, three funnel plots, trim-and-fill plot, Galbraith plot, and L’Abbé plot) and assessed the resulting 1100 plots with three reviewers. The intra-class correlation coefficients for reproducibility of the graph ratings ranged from poor (0.34) to high (0.91). Ratings of the forest plot and the standardized residual histogram were best associated with parameter heterogeneity. Association between graph ratings and dissemination selectivity (censorship of studies) was poor, with the weighted box plot and the trim-and-fill plot showing the best combination of reproducibility and validity. Meta-analysts should therefore be selective in the graphs they choose for the exploration of their data.

Chapter 5 describes the development of meta-analysis software. Because there was a niche for a cheap, easy-to-use, and comprehensive meta-analysis package, we started a meta-analysis software development project. An important part of the development focused on creating internal interactive tutoring tools, and on designing features that would facilitate usage of the software as a companion to existing books on meta-analysis. We took an unconventional approach and created a program that uses Excel as a calculation and programming platform. The main programming language was Visual Basic, as implemented in Visual Basic 6 and Visual Basic for Applications in Excel 2000 and higher. The development took approximately two years and resulted in the MIX program. Next, we set out to validate the MIX output with two major software packages as reference standards, namely STATA (metan, metabias, and metatrim) and Comprehensive Meta-Analysis (CMA). Eight meta-analyses that had been published in major journals were used as data sources. All numerical and graphical results from analyses with MIX were identical to their counterparts in STATA and CMA. The MIX program distinguishes itself from most other programs by the extensive graphical output,

the click-and-go (Excel) interface, and the educational features. The MIX program is a valid tool for performing meta-analysis and may be particularly useful in educational environments. It can be downloaded free of charge via http://www.mix-for-meta-analysis.info.

In **Chapter 6**, we report a systematic review of the differences in features, results, and usability of currently available meta-analysis programs. We did an extensive search on the Internet (Google, Yahoo, Altavista, and MSN) for specialized meta-analysis software, and eventually included six programs in our review: Comprehensive Meta-Analysis (CMA), MetAnalysis, MetaWin, MIX, RevMan, and WEasyMA. Two investigators compared the features of the software and their results. Thirty independent researchers evaluated the programs on their usability while analyzing one data set. The programs differed substantially in features, ease-of-use, and price. Although most results from the programs were identical, we did find some minor numerical inconsistencies. CMA and MIX scored the highest on usability and these programs also have the most complete set of analytical features. Considering the differences in numerical results, we believe the user community would benefit from openly available and systematically updated information about the procedures and results of each program's validation. The most suitable program for a meta-analysis will depend on the user's needs and preferences, and this chapter provides an overview that should be helpful in making an informed choice.

10.3 Applications

The applications part contains two actual meta-analyses applying the methods and software discussed in the previous part. **Chapter 7** contains a re-analysis of a recent, widely publicized meta-analysis of 42 clinical trials which concluded that rosiglitazone was associated with an approximately 43% increased risk for myocardial infarction and an approximately 64% increased risk for cardiovascular death. We felt that the original analysis was inadequate as the sensitivity of the conclusions to several methodological choices was not assessed. Moreover, studies were combined despite substantial variability in study design and outcome assessment, and the meta-analytic approach that was used required the exclusion of many studies with zero events in the treatment and control groups. In our meta-analysis, we show that alternative meta-analytic approaches consistently yield results that are not statistically significant. We conclude that the risk for myocardial infarction and death from cardiovascular disease for diabetic patients taking rosiglitazone is uncertain: neither increased nor decreased risk is established.

In **Chapter 8** we report a systematic review and meta-analysis of randomized controlled trials to determine whether neuromuscular electrical stimulation (NMES) is an effective modality for strength augmentation of the quadriceps femoris. A full content search for randomized controlled trials was performed in Medline, Embase, Cinahl, the Cochrane Controlled Trials Register, and the Physical Therapy Evidence Database. Maximum volitional isometric or isokinetic muscle torque in Nm was used as main outcome measure. Thirty-five trials were included and evaluated. The included trials were generally of poor quality and meta-analytic data indicate that publication bias may be present. The evaluated data suggest that, both for the unimpaired and impaired quadriceps, NMES makes sense compared with doing no exercises but volitional exercises appear to be more effective in most situations. NMES may only be preferable over volitional training for within-cast muscle training and perhaps in specific situations where volitional training does not receive sufficient patient compliance. Further research should be directed toward identifying the clinical impact at activity and participation levels and the optimal stimulation parameters of this modality.

10.4 Discussion and summary

Instead of a step-by-step discussion of the contents of the previous chapters, **Chapter 9** provides a comprehensive review of methods for causal meta-analysis, including the topics discussed in this dissertation. It provides a practical guide for conducting causal meta-analysis with a clear distinction between the exploration, synthesis, and evaluation stages. Instructions and syntax for the statistical software packages MIX, R, and OpenBUGS are provided to facilitate the application of the proposed methods. We stress that meta-analysis is more than a method to establish a (significant) pooled result based on aggregate-level data from multiple studies. Meta-analysts should routinely go beyond the synthesis, explore their data thoroughly, and evaluate the sensitivity of results to alternative statistical approaches.

11

SAMENVATTING IN HET NEDERLANDS

11.1 Introductie

In **Hoofdstuk 1** introduceren we meta-analyse aan de hand van enkele illustratieve voorbeelden. We benadrukken dat een meta-analyse onderdeel behoort te zijn van een systematische review. Medisch onderzoek kan benaderd worden vanuit een causaal of descriptief perspectief en we beschrijven hoe dit proefschrift zich met name bezighoudt met meta-analyses van causale onderzoeksdata. Meta-analyses zijn niet altijd mogelijk, zoals bijvoorbeeld wanneer er grote verschillen bestaan tussen de studies die met elkaar gecombineerd zouden worden in de meta-analyse. Hedendaags zijn meta-analyses echter onvervangbaar bij het tot stand komen van klinische beslissingsmodellen en medische richtlijnen.

11.2 Methodologie

Het tweede gedeelte van dit proefschrift bevat een aantal methodologische artikelen over causale meta-analyse. In **Hoofdstuk 2** geven we een eenvoudig overzicht van wie meta-analyses publiceert. Vaak wordt aangenomen dat meta-analyses worden gedaan door doorgewinterde onderzoekers, maar dit blijkt niet altijd het geval te zijn. Op basis van een zoekactie in Medline (PubMed) vinden we dat in 2008 ongeveer 59% van de eerste auteurs van meta-analyses niet eerder aan een meta-analyse hebben meegewerkt. Slechts 28% heeft eerder als eerste auteur een meta-analyse gepubliceerd. De tweede en laatste auteurs waren meer ervaren. Kortom, meta-analisten zijn niet altijd experts op meta-analyse gebied. Dit onderstreept het belang van goede peer-reviews en heeft gevolgen voor de vormgeving van educatief materiaal en software voor meta-analisten.

Hoofdstuk 3 bevat een korte verhandeling over publicatie bias. Deze term wordt vaak gebruikt om te beschrijven hoe onderzoekers, reviewers, en redacteurs vooral (statistisch) significante en positieve resultaten publiceren. Een aantal methodologische publicaties geven aan dat de term ongelukkig en onnauwkeurig is, maar er is geen consensus over alternatieve terminologie. We bespreken de gebruikte termen in detail en doen voorstellen voor alternatieven. We onderscheiden selectiviteit in rapportage (van zowel uitkomsten in een studie als studies in het geheel), publicatie, en inclusie, en stellen dat deze selectiviteit een systematische vertekening van resultaten (bias) tot gevolg kan hebben. Binnen deze benadering zijn rapportage bias, publicatie bias, en inclusie bias onderdelen van disseminatie bias. Disseminatie bias verwijst naar systematische vertekening ten gevolge van een selectiviteit in de loop van de openbaarmaking van onderzoeksgegevens. Het is duidelijk dat in ons model publicatie bias een ‘pars pro toto’ is (een onderdeel om het geheel te beschrijven) en daarom niet gebruikt dient te worden als paraplu-term. In hetzelfde kader maken we onderscheid tussen

statistische methodes die trachten vast te stellen of er selectiviteit heeft plaatsgevonden en methodes die trachten te kwantificeren in hoeverre de selectiviteit leidt tot een vertekening van de resultaten van de meta-analyse.

Grafieken worden veelvuldig gebruikt in meta-analyses om te beoordelen in hoeverre heterogeniteit en disseminatie-selectiviteit een rol hebben gespeeld in de totstandkoming van de data set. Een verscheidenheid aan grafieken worden gebruikt, maar tot op heden heeft er geen vergelijkende evaluatie plaatsgevonden van de prestaties van de grafieken. In **Hoofdstuk 3** beoordelen we de reproduceerbaarheid en validiteit van de beoordelingen van grafieken. Op basis van vier scenario's (met verschillende mate van heterogeniteit en selectiviteit) werden 100 meta-analyses gesimuleerd. Voor iedere meta-analyse data set werden 11 grafieken gecreëerd (box plot, weighted box plot, standardized residual histogram, normal quantile plot, forest plot, drie funnel plots, trim-and-fill plot, Galbraith plot, en L'Abbé plot) en de grafieken werden beoordeeld door drie reviewers. De intra-class correlatie-coëfficiënten van de reproduceerbaarheid van de grafiek beoordelingen liepen uiteen van 0.34 tot 0.91. De beoordelingen van de forest plot en het standardized residual histogram kwamen het best overeen met de gesimuleerde heterogeniteit. De associatie tussen de beoordelingen van de grafieken en de gesimuleerde selectiviteit was laag; de weighted box plot en de trim-and-fill plot presteerden hier het best. Samenvattend kan gesteld worden dat onderzoekers die betrokken zijn bij meta-analyses selectief dienen te zijn bij de keuze van de grafieken die gebruikt worden in de meta-analyse.

Hoofdstuk 5 beschrijft de ontwikkeling van meta-analyse software. Bij gebrek aan een goedkoop en tegelijkertijd veelomvattend en gebruiksvriendelijk software pakket voor meta-analyse initieerden we een software ontwikkelingsproject. Een belangrijk onderdeel van het project was gericht op de ontwikkeling van ingebouwde tutorials en software onderdelen die het gebruik van de software naast bestaande meta-analyse boeken zouden vergemakkelijken. We namen een ongewone benadering en gebruikten Microsoft Excel als platform voor het programmeren en maakten een uitgebreide Excel add-in met Visual Basic als programmeertaal. De ontwikkeling duurde ongeveer twee jaar met als resultaat het MIX programma. De validatie op basis van acht meta-analyse data sets met de twee veelgebruikte meta-analyse paketten STATA (metan, metabias, en metatrim) en Comprehensive Meta-Analysis (CMA) liet zien dat alle numerieke resultaten en grafieken van MIX identiek zijn aan de tegenhangers in STATA and CMA. Het MIX programma onderscheidt zich van de andere programma's door een uitgebreid scala aan grafieken, de Excel gebruikersinterface, en de educatieve onderdelen. MIX is een valide statistisch pakket voor meta-analyse en vooral ook

geschikt voor meta-analyse scholing. Het is gratis te downloaden via http://www.mix-for-meta-analysis.info.

In **Hoofdstuk 6** wordt een systematische review van meta-analyse software beschreven. Er wordt onder meer gekeken naar de statistische mogelijkheden, validiteit, en gebruiksvriendelijkheid van de beschikbare meta-analyse pakketten. Via het internet (Google, Yahoo, Altavista, en MSN) werd gezocht naar meta-analyse software en uiteindelijk werden zes software pakketten geïncludeerd in de review: Comprehensive Meta-Analysis (CMA), MetAnalysis, MetaWin, MIX, RevMan, and WEasyMA. Twee onderzoekers vergeleken de statistische mogelijkheden en de resultaten van analyses met drie data sets. Daarnaast evalueerden dertig onafhankelijke onderzoekers de software op gebruiksvriendelijkheid. De programma's toonden aanzienlijke verschillen met betrekking tot mogelijkheden, gebruiksvriendelijkheid, en prijs. De meeste resultaten van analyses met de programma's waren identiek. CMA and MIX scoorden het hoogst wat betreft gebruiksvriendelijkheid en deze programma's bieden ook de meeste analytische vrijheid. Het meest geschikte programma voor het uitvoeren van een meta-analyse hangt af van de eisen en voorkeuren van de gebruiker.

11.3 Toepassingen

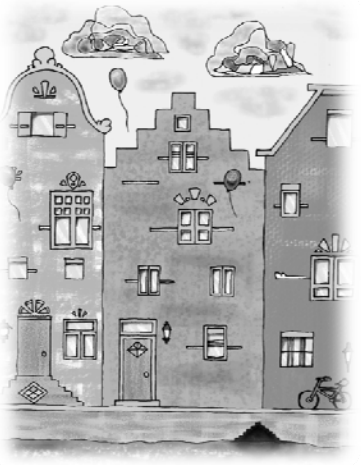
Hoofdstuk 7 beschrijft een her-analyse van de onderzoeksdata van een recente, veelbesproken meta-analyse van 42 klinische onderzoeken naar de veiligheid van rosiglitazone. De oorspronkelijke meta-analyse concludeerde dat rosiglitazone mogelijk verantwoordelijk was voor een met 43% verhoogd risico op een hart-infarct en een met 64% verhoogd risico op overlijden ten gevolge van een cardiovasculaire aandoening. De oorspronkelijke meta-analyse presenteerde echter geen gegevens over hoe stabiel deze resultaten waren met betrekking tot de keuze van de meta-analytische methodes. Tevens werden alle studies gecombineerd in de synthese, ondanks het feit dat er substantiële verschillen waren in de opzet van de studies en de meetwijze van de uitkomstmaten. Tegelijkertijd werden studies zonder hart-infarcten of sterfgevallen niet effectief betrokken in de statistische synthese. In onze analyse tonen we aan dat de resultaten wel degelijk gevoelig zijn voor de gebruikte meta-analytische methodes en dat alle alternatieve benaderingen consequent leiden tot resultaten die niet statistisch significant zijn. Wij concluderen daarom dat het risico op hartfalen en cardiovasculaire sterfte voor diabetes-patiënten die rosiglitazone gebruiken dient te worden geclassificeerd als onzeker: op basis van de beschikbare data kan noch een verhoogd, noch een verlaagd risico worden aangetoond.

In **Hoofdstuk 8** wordt beschreven in hoeverre neuromusculaire elektrische stimulatie (NMES) effectief is in het versterken van de quadriceps femoris spiergroep. Gerandomiseerde klinische studies werden gezocht in Medline, Embase, Cinahl, het Cochrane Controlled Trials Register, en de Physical Therapy Evidence Database. De maximale willekeurige isometrische of isokinetische kracht in Nm werd gebruikt als primaire uitkomstmaat. Vijfendertig studies werden geïncludeerd en geëvalueerd in verschillende subgroepen. De studies waren over het algemeen van lage kwaliteit. De resultaten van de synthese tonen aan dat NMES tot spierversterking kan leiden, maar dat actieve spiertraining betere resultaten geeft. NMES kan echter ook gebruikt worden bij patiënten waarbij de knie is geïmmobiliseerd en actieve oefeningen niet of niet op een gecontroleerde wijze uitgevoerd kunnen worden. Meer onderzoek is nodig om de optimale parameters voor spierstimulatie te bepalen.

11.4 Discussie en samenvatting

Hoofdstuk 9 geeft een uitgebreid overzicht van moderne methodes voor causale meta-analyse. Het onderscheidt drie fases in een meta-analyse: exploratie van de ruwe data, synthese van de resultaten van de studies, en een evaluatie van de stabiliteit van de synthese ten aanzien van het gebruik van alternatieve statistische benaderingen. Instructies en syntax voor de statistische software MIX, R, en WinBUGS faciliteren de toepassing van de beschreven methodiek. De onderwerpen variëren van eenvoudige fixed-effect modellen tot Bayesiaanse hiërarchische regressie-analyses. We benadrukken dat een meta-analyse meer is dan het combineren van data van studies om te komen tot een (significant) samenvattend resultaat. Onderzoekers die betrokken zijn bij meta-analyses hebben de verantwoordelijkheid verder te gaan dan deze synthese. De exploratie van ruwe data alsook de evaluatie van de stabiliteit van de synthese resultaten ten aanzien van alternatieve methodes horen thuis in iedere meta-analyse.

APPENDICES



A

GETTING STARTED WITH MIX

A.1 Acquiring MIX

The easiest way to get the most recent version of MIX is to download it via the MIX Web site at *http://www.mix-for-meta-analysis.info*. There are a substantial number of links to the site from Web sites about meta-analysis and statistics software, so typing "MIX" and "meta-analysis" in any search engine should get you to the right place as well (Figure A.1).

Figure A.1 MIX home page

The MIX home page with navigation menu at the top and general information about the program and the Web site with links in the central part.

A.2 Installing MIX

You should install MIX on a computer with a processor that is equivalent or superior to an Intel Pentium ® 400 MHz (800 MHz or higher recommended), 256 MB of RAM (512 MB or higher recommended), 30 MB of available disk space (version 1.7 of the program will take up about 20 MB), and a color monitor capable of 800 x 600 resolution (1074 x 768 recommended).

As far as software is concerned, MIX requires a computer with Microsoft Windows and Microsoft Excel 2000 or later. Users of Excel 2007 are recommended to use MIX 2.0 and users of earlier versions of Excel will need to use MIX 1.7. Finally, it is important that you have installed the most recent

updates for Microsoft Excel and Microsoft Office. Please note that Windows Updates are not the same as Office updates.

For some reading materials a PDF reader is required (Adobe Acrobat Reader can be downloaded from the Internet for free), and for the Program Tour and Theory Tour you need a web browser with Flash plug-in (we recommend Mozilla Firefox).

Once you have made sure your system fulfills the requirements, you can double-click the icon of the set-up program that you have downloaded from the MIX website. The installation program will start and if you follow the instructions, you will soon have a new menu in your Start Folder called MIX. Depending on your choices and MIX version, you may also have a new MIX icon on your Desktop or a MIX menu in your Excel ribbon. You are now ready to use MIX.

A.3 Using MIX

It goes beyond a short appendix to provide a step-by-step guide to using MIX. However, the MIX Web site is regularly updated with tutorials (PDF and interactive Flash ®) at *http://www.mix-for-meta-analysis.info/about/tutorial.html*.

MIX was designed to be accessible to researchers without much exposure to meta-analysis and the first use should be as intuitive as it can be. We recommend starting with the built-in data sets that can be found in the Data Wizard. The MIX Web site also contains a number of tutorials to show you how to get started. The built-in Output Tutor can also provide you with advice and information about the analysis you are performing.

MIX 1.7 is free for all purposes and works with Excel versions up to Office 2003. The new version of MIX (MIX 2.0) is compatible with Excel 2007. However, MIX 2.0 has some limitations for non-registered users. Everybody can still use the program for educational purposes and there are no restrictions on the analyses that can be done on the built-in data sets. However, if you would like to use MIX 2.0 for analysis of your own data, you will have to register as a user. Students and users in developing countries can register for free, but academic, private, and commercial users will have to pay a small fee. See the *MIX Web site* for details.

B

CODE FOR R AND OPENBUGS

B.1 Load the data set

```
# -----------------
# File: load data.r
# -----------------
# Set the working directory
# Default is the directory of the MIX software
  setwd('C:/Program Files/MIX/Resources/R')
# Get the data set
  dset=read.table("mi.csv", header = TRUE, sep = ",")
```

B.2 Data set exploration

```
# ----------------------------
# File: data set exploration.r
# ----------------------------
# Look at the data set first
  dset # check its format
  nstudies=length(dset$id)
  nstudies
  ssize=dset$g1.n + dset$g0.n # create sample size variable
  summary(dset) # data set summary
# Some ways to look at the sample sizes
  ssize
  summary(ssize)
  boxplot(ssize)
  hist(ssize,freq=TRUE,xlim=c(0,10000),main="Sample size histogram",
  breaks=200)
  stripchart(ssize, method="stack", pch=1, offset=0.5, cex=2)
  dset$g1.r1
  dset$g0.r1
  summary(dset$g1.r1)
  summary(dset$g0.r1)
  stripchart(dset$g1.r1, method="stack", pch=1, offset=0.5, cex=2)
  stripchart(dset$g1.r1, method="stack", pch=1, offset=0.5, cex=2)
  stem(dset$g1.r1, scale =3)
  stem(dset$g0.r1, scale =3)
# Check explicitly for zero events
  attach(dset)
  g1zero=sum(g1.r1==0&g0.r1!=0) # indicator of zero events group 1
  g0zero=sum(g0.r1==0&g1.r1!=0) # indicator of zero events group 0
  gzerosingle=g1zero+g0zero
  gzerosingle
  gzeroall=sum(g1.r1==0&g0.r1==0) # indicator of total zero events
  gzeroall
  detach(dset)
```

B.3 Graphical model exploration

```
# ----------------------------------
# File: graphical model exploration.r
# ----------------------------------
# Load the meta package first
# Install first if necessary (uncomment the install lines)
# install.packages("meta")
  library(meta)
# A quick fixed effect synthesis to create some variables
  meta1=metabin(g1.r1,g1.n,g0.r1,g0.n,data=dset,sm="OR",method="I")
```

```
# Forest plot
  plot.meta(meta1,studlab=FALSE, main="Forest plot")
# Standardized residual histogram with normal overlay
  st.res=(meta1$TE-meta1$TE.fixed)/meta1$seTE
  hist(st.res,freq=FALSE,ylim=c(0,1),main="Standardized residual
  histogram", breaks=20)
  curve(dnorm(x,mean(st.res),sd(st.res)),add=TRUE,lty=4)
# L'Abbe plot
  g1.o1=(g1.r1/g1.n)/(1-(g1.r1/g1.n))
  g0.o1=(g0.r1/g0.n)/(1-(g0.r1/g0.n))
  g1.p1=(g1.r1/g1.n)
  g0.p1=(g0.r1/g0.n)
  plot(g1.p1~g0.p1,main="L'Abbe plot", xlim=c(0,0.5), ylim=c(0,0.5),
  cex=15*meta1$w.fixed/max(meta1$w.fixed))
  res=lm(g1.p1~g0.p1)
  abline(a=0,b=coef(res))
  abline(0,1)
# Galbraith plot
  radial(meta1,level=0.95,main="Galbraith plot")
```

B.4 Numerical model exploration

```
# ----------------------------------
# File: numerical model exploration.r
# ----------------------------------
# Load the meta package first
# Install first if necessary (uncomment the install lines)
# install.packages("meta")
  library(meta)
# A quick fixed effect synthesis to create the heterogeneity variables
  meta1=metabin(g1.r1,g1.n,g0.r1,g0.n,data=dset,sm="OR",method="I")
# Get the Q, I-square, and tau-square statistics
  het.output=function(meta)
    {
    cat("-----------------------------------","\n")
    cat("Testing the fixed effect assumption","\n")
    cat("-----------------------------------","\n")
    cat("Q-statistic = ",meta$Q,sep="")
    cat(" (df = ",meta$k-1, sep="")
    cat(", P = ",pchisq(meta1$Q,df=meta1$k-1, lower.tail=FALSE),
    ")",sep="", "\n")
    cat("I-squared = ",max(0,((meta$Q-(meta$k-1))/meta$Q)*100),sep="")
    cat("%), sep="","\n")
    cat("Tau-squared =",meta$tau^2, fill=TRUE)
    }
  het.output(meta1)
  meta1
```

B.5 Traditional synthesis

```
# ----------------------------
# File: traditional synthesis.r
# ----------------------------
# Load the meta package first
# Install first if necessary (uncomment the install lines)
# install.packages("meta")
  library(meta)
# An example of an OR synthesis from event data
  # Mantel-Haenszel
```

```
  meta.or.mh=metabin(g1.r1,g1.n,g0.r1,g0.n, data=dset, sm="OR",
  method="MH")
  summary(meta.or.mh)
# Inverse variance
  meta.or.iv=metabin(g1.r1,g1.n,g0.r1,g0.n,data=dset,sm="OR",
  method="I")
  summary(meta.or.iv)
# Peto
  meta.or.p=metabin(g1.r1,g1.n,g0.r1,g0.n,data=dset,sm="OR",
  method="P")
  summary(meta.or.p)
# Overview of weights
  w.or.iv=round(meta.or.iv$w.fixed/sum(meta.or.iv$w.fixed),2)
  w.or.mh=round(meta.or.mh$w.fixed/sum(meta.or.mh$w.fixed),2)
  w.or.p=round(meta.or.p$w.fixed/sum(meta.or.p$w.fixed),2)
  w.or.dl=round(meta.or.iv$w.random/sum(meta.or.iv$w.random),2)
  prop.weights=cbind(w.or.iv, w.or.mh, w.or.p, w.or.dl)
  prop.weights
# Overview of weights as percentages
  w.or.iv.tx=paste(round(meta.or.iv$w.fixed/sum(meta.or.iv$w.fixed)
  *100,2),"%",sep="")
  w.or.mh.tx=paste(round(meta.or.mh$w.fixed/sum(meta.or.mh$w.fixed)
  *100,2),"%",sep="")
  w.or.p.tx=paste(round(meta.or.p$w.fixed/sum(meta.or.p$w.fixed)
  *100,2),"%",sep="")
  w.or.dl.tx=paste(round(meta.or.iv$w.random/sum(meta.or.iv$w.random)
  *100,2),"%",sep="")
  prop.weights.tx=cbind(w.or.iv.tx, w.or.mh.tx, w.or.p.tx, w.or.dl.tx)
  prop.weights.tx
```

B.6 Fixed effect linear regression

```
# --------------------------------------
# File: fixed effect linear regression.r
# --------------------------------------
# Load the meta package first
# Install first if necessary (uncomment the install lines)
# install.packages("meta")
  library(meta)
# First do standard inverse variance synthesis and get zero cells
  meta1=metabin(g1.r1,g1.n,g0.r1,g0.n,data=dset,sm="OR",method="I")
  attach(dset)
  g1zero=sum(g1.r1==0&g0.r1!=0) # indicator of zero events only in one
  group 1
  g0zero=sum(g0.r1==0&g1.r1!=0) # indicator of zero events only in
  group 0
  gzeroall=sum(g1.r1==0&g0.r1==0) # indicator of total zero events
  detach(dset)
# Prepare regression input
  est.i=meta1$TE  # individual 'normal' estimates
  w.iv=meta1$w.fixed # weights
  var.i=meta1$seTE^2 # variances
# Regression on only the intercept
  res=lm(est.i~1,weight=w.iv) # regression on the intercept only
  n=length(est.i)-gzeroall # number of studies
  ncoef=length(coef(res)) # number of coefficients (just 1)
  rss=deviance(res) # residual sum of squares
  rms=rss/(n-ncoef) # residual mean squares
  log.est.sum=coef(summary(res))[1,1] # log of summary estimate
  est.sum=exp(log.est.sum) # estimate
```

```
  se=coef(summary(res))[ 1,2]/sqrt(rms) # adjusted standard error
  #distdev=qt(1-0.05/2,df=n-2) # t deviate
  distdev=qnorm(1-0.05/2) # z deviate
  logci=c(log.est.sum-se*distdev,log.est.sum+se*distdev) # confidence
  interval of log-transformed estimate
  logci
  ci=exp(logci) # confidence interval of estimate
  c(est.sum,ci) # summary
```

B.7 Random effects linear regression

```
# ---------------------------------------
# File: random effects linear regression.r
# ---------------------------------------
# Load the meta package first
# Install first if necessary (uncomment the install lines)
# install.packages("meta")
  library(meta)
# First do standard inverse variance synthesis and get zero cells
  meta1=metabin(g1.r1,g1.n,g0.r1,g0.n,data=dset,sm="OR",method="I")
  attach(dset)
  g1zero=sum(g1.r1==0&g0.r1!=0) # indicator of zero events only in one
  group 1
  g0zero=sum(g0.r1==0&g1.r1!=0) # indicator of zero events only in
  group 0
  gzeroall=sum(g1.r1==0&g0.r1==0) # indicator of total zero events
  detach(dset)
# Prepare regression input
  est.i=meta1$TE  # individual 'normal' estimates
  w.iv=meta1$w.fixed # weights
  var.i=meta1$seTE^2 # variances
  est.i=na.omit(est.i)
  w.iv=na.omit(w.iv)
  var.i=na.omit(var.i)
# Random effects regression on only the intercept
  source("mima.ssc")
  mods=c()
  mima(est.i,var.i,mods,method="DL",fe="no",out="no")
  # Exponentiate the intercept and the confidence limits
  exp(-0.4331)
  exp(-0.6304)
  exp(-0.2357)
  # Use a different estimator
  mima(est.i,var.i,mods,method="REML",fe="no",out="no")
```

B.8 Logistic contingency table regression

```
# ---------------------------------------------
# File: logistic contingency table regression.r
# ---------------------------------------------
# Get the contingency table data set
  dset=read.table("micontingency.csv", header = TRUE, sep = ",")
# Prepare the data
  y.i=dset$r ; x1.i=dset$g ; w.i=dset$subjects
# Logistic regression in glm
  res = glm(y.i~x1.i,family=binomial(),weights=w.i)
  summary(res)
  exp(coef(res)); exp(confint(res))
```

B.9 Logistic individual regression

```
# -------------------------------------
# File: logistic individual regression.r
# -------------------------------------
# Get the data set
  dset=read.table("miindividual.csv", header = TRUE, sep = ",")
# Prepare the data
  y.i=dset$r ; x1.i=dset$g
  level.j=dset$studynum
# Hierarchical logistic regression
  res = lmer(y.i~x1.i+(1|level.j),family=binomial(link=logit))
  summary(res)
  coef(res); exp(fixef(res)[2])
# Calculate the CIs by hand with the reported standard error in the
  summary
```

B.10 Bayesian hierarchical binomial model for OpenBUGS

```
# -------------------------------------
# File: bayesian hierarchical binomial.r
# -------------------------------------
model{
# Set up a model with k plates (repeated structures) to represent the
  studies
    for(i in 1:k){
        # likelihoods for events in exposed and control group
          g1.r1[i]~dbin(g1.p[i],g1.n[i])
          g0.r1[i]~dbin(g1.p[i],g0.n[i])
        # define the effects
          logit(g0.p[i])<-bm[i]; logit(g1.p[i])<-bm[i]+tm[i]
        # tm and bm can be modeled as normal
          bm[i]~dnorm(0.0,1.0E-5)
          tm[i]~dnorm(mu.tm,tau.tm)
        # produce individual odds ratios for graph
          OR[i]<-exp(logit(g1.p[i])-logit(g0.p[i]))
    }
    # set a normal diffuse prior distribution for the treatment location
      parameter in the random effects setting
      mu.tm~dnorm(0.0,1.0E-5)
    # give the precision a diffuse gamma prior distribution in the
      random effects setting
      tau.tm~dgamma(0.001,0.001)
    # transform back to the linear scale to get an odds ratio
      OR.tm<-exp(mu.tm)
}
```

B.11 Random effects linear regression with covariate

```
# -------------------------------------
# File: linear random effects covariate.r
# -------------------------------------
# Load the meta package first
# Install first if necessary (uncomment the install lines)
# install.packages("meta")
  library(meta)
# Get the data set again
  dset=read.table("mi.csv", header = TRUE, sep = ",")
```

```
# First do standard inverse variance synthesis and get zero event counts
  meta1=metabin(g1.r1,g1.n,g0.r1,g0.n,data=dset,sm="OR",method="I")
  attach(dset)
  g1zero=sum(g1.r1==0&g0.r1!=0) # indicator of zero events only in one
  group 1
  g0zero=sum(g0.r1==0&g1.r1!=0) # indicator of zero events only in
  group 0
  gzeroall=sum(g1.r1==0&g0.r1==0) # indicator of total zero events
  detach(dset)
# Prepare regression input
  est.i=meta1$TE  # individual 'normal' estimates
  w.iv=meta1$w.fixed # weights
  var.i=meta1$seTE^2 # variances
  est.i=na.omit(est.i)
  w.iv=na.omit(w.iv)
  var.i=na.omit(var.i)
# Random effects regression with covariate
  source("mima.ssc")
  # specify 14 covariates
  mods=c(6,6,7,8,4,6,7,5,6,8,9,4,6,8)
  # run the model
  mima(est.i,var.i,mods,method="DL",fe="no",out="no")
  # exponentiate the intercept and the confidence limits manually
```

B.12 Funnel plots

```
# --------------------
# File: funnel plots.r
# --------------------
# Load the meta package first
# Install first if necessary (uncomment the install lines)
# install.packages("meta")
  library(meta)
# Get the data set
  dset=read.table("mi.csv", header = TRUE, sep = ",")
# First do standard inverse variance synthesis
  meta1=metabin(g1.r1,g1.n,g0.r1,g0.n,data=dset,sm="OR",method="I")
# Funnel plot with standard error on y-axis and pseudo-confidence limits
  funnel(meta1, comb.f=TRUE, level=0.95)
# Funnel plot with sample size on y-axis
  funnel(meta1,comb.f=TRUE,yaxis="size")
# Funnel plot with inv standard error on y-axis
  funnel(meta1,comb.f=TRUE,yaxis="invse")
```

B.13 Bias tests

```
# --------------------
# File: bias tests.r
# --------------------
# Load the meta package first
# Install first if necessary (uncomment the install lines)
# install.packages("meta")
  library(meta)
# First do standard inverse variance synthesis
  meta1=metabin(g1.r1,g1.n,g0.r1,g0.n,data=dset,sm="OR",method="I")
# Tests for small-study effects
  metabias(meta1,method="rank") # with rank correlation
  metabias(meta1,method="linreg", plotit=TRUE) # Egger's regression
```

B.14 Trim and fill assessment

```
# --------------------
# File: trim and fill.r
# --------------------
# Load the meta package first
# Install first if necessary (uncomment the install lines)
# install.packages("meta")
  library(meta)
# First do standard inverse variance synthesis
  meta1=metabin(g1.r1,g1.n,g0.r1,g0.n,data=dset,sm="OR",method="I")
# Perform trim-and-fill analysis
  trimfill(meta1,type="L")
```

ACKNOWLEDGMENTS

This dissertation is the product of the efforts of many. I would therefore like to start by expressing my gratitude to everybody who has supported me in my academic and private life over the past few years.

Next, I would like to thank Professor Carl Moons, my supervisor at the Julius Center for Health Sciences and Primary Care. Carl, your ideas, critical comments, and continuous support have been invaluable. I would also like to thank my former supervisor at Kitasato University in Japan, Professor Noriaki Ikeda, and my current boss at the Kitasato Clinical Research Center, Professor Toshihiko Satoh. You have both been a great help. Many thanks also to Harukazu Tsuruta, who has been a companion and guide in topics ranging from mathematics to Japanese language and culture. I am very grateful to Shizuka Tsuruta, who made time between her exams to design the cover of this dissertation.

I would like to express my gratitude to the other co-authors of the papers included in this dissertation: Ly-Mee Yu, George Diamond, Sanjay Kaul, Arianne Verhagen, Filip Staes, Yukari Yaju, Naohito Fukui, and Guoqin Wang. A special thank you goes out to Akihiro Takeuchi, Masuo Shirataka, Yasuo Morohoshi, Sadayasu Shibata, Osamu Henmi, Noritaka Mamorita, and everybody of the Department of Medical Informatics at Kitasato University for their comments and advice over the years.

I would also like to thank Carl's secretary, Annina Koopmans. Completing the doctoral program in the Netherlands while living and working in Japan was not easy, and you have been a great help on the Dutch side. Thanks also to Kristel Janssen for giving me some inside information while preparing the dissertation and the defense.

I could not have done my work in Japan without the support of my family in the Netherlands. Life has been tough, losing mom in 2008, but we are all doing our best to get on with our lives. I think she would have been proud. I would also like to extend a big thank you to my family-in-law. You have been a great support for me and Nina.

Speaking of Nina… "Ma petite …!!!" We have gone through a lot over the past few years. You have not only been an incredible support during my research and the writing of this dissertation, but life is just so much fun together!

About the author

I grew up in the Netherlands where I attended the Gymnasium high school in Gorinchem from 1985 to 1991. I subsequently went to the University of Texas in San Antonio (the United States) on an athletic scholarship. Although playing tennis was fun, it didn't give me the satisfaction I was looking for. I returned to the Netherlands in 1992, and enrolled in the Physical Therapy program of the Utrecht University of Applied Sciences, while working as a tennis coach. I graduated in 1996 and worked as a physical therapist at a number of institutions. Somewhat disappointed by the content of my work, I decided to look for another challenge and enrolled in the Japanese Language and Culture graduate program at the University of Leiden (the Netherlands). After spending some time in Japan on a scholarship, I returned to Europe. While working in various capacities, I completed two Master of Science programs: Physical Therapy at the University of Leuven in Belgium in 2003, and Clinical Epidemiology at the Netherlands Institute for Health Sciences in the Netherlands in 2004. My wife and I moved to Japan in 2003, where I initially worked and studied at Kitasato University as a government-funded research student in Medical Informatics. I later enrolled in the doctoral program in Medical Informatics, which I completed in early 2008. The current dissertation concludes my doctoral program in Clinical Epidemiology at the Julius Center for Health Sciences and Primary Care in Utrecht, The Netherlands. At the time of writing this manuscript, we still live in Japan and I am employed as a Junior Associate Professor in Clinical Epidemiology & Biostatistics at the Kitasato University School of Medicine and the Kitasato Clinical Research Center. Although this dissertation primarily reports some of the methodological research I have done, I am increasingly involved in primary studies via the Kitasato Clinical Research Center and through consulting work for the pharmaceutical industry.

Meta-analysis is a statistical method to summarize research data from multiple studies in a quantitative manner. This dissertation addresses a number of methodological topics in causal meta-analysis and reports the development and validation of meta-analysis software. In the first (methodological) part, we explore how experienced or inexperienced authors of meta-analyses are, discuss the terminology related to dissemination bias, explore the usefulness of graphical assessments, and report the development of software (MIX) for causal meta-analysis. This is followed by an applications part with two actual meta-analyses. In this part, the previously discussed methods are applied and the MIX software is used. The first meta-analysis re-analyzes data from a recent, widely publicized meta-analysis of 42 clinical trials which concluded that rosiglitazone was associated with an approximately 43% increased risk for myocardial infarction and an approximately 64% increased risk for cardiovascular death. The other meta-analysis explores the effects of neuromuscular electrical stimulation on muscle strength. The dissertation is concluded with a practical methodological paper on modern methods of causal meta-analysis, providing a step-by-step guide on how to apply a wide range of meta-analytical methods to causal research data.