10-1-2011

# Contributions to estimation and interpretation of intervention effects and heterogeneity in meta-analysis

Kristian Thorlund
*McMaster University*, thorluk@mcmaster.ca

# CONTRIBUTIONS TO ESTIMATION AND INTERPRETATION OF INTERVENTION EFFECTS AND HETEROGENEITY IN META-ANALYSIS

BY

KRISTIAN THORLUND, B.Sc., M.Sc.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

Doctor of Philosophy

McMaster University

DOCTOR OF PHILOSOPHY 2011 McMaster University

(Health Research Methodology) Hamilton, Ontario

TITLE: CONTRIBUTIONS TO ESTIMATION AND INTERPRETATION OF INTERVENTION EFFECTS AND HETEROGENEITY IN META-ANALYSIS

AUTHOR: KRISTIAN THORLUND, B.Sc. (University of Copenhagen), M.Sc. (University of Copenhagen)

SUPERVISORS: Dr. Lehana Thabane, Dr. Philip James Devereaux

NUMBER OF PAGES: xiv, 214

## Abstract

### *Background and objectives*

Despite great statistical advances in meta-analysis methodology, most published meta-analyses make use of out-dated statistical methods and authors are unaware of the shortcomings associated with the widely employed methods. There is a need for statistical contributions to meta-analysis where: 1) improvements to current statistical practice in meta-analysis are conveyed at the level that most systematic review authors will be able to understand; and where: 2) current statistical methods that are widely applied in meta-analytic practice undergo thorough testing and examination. The objective of this thesis is to address some of this demand.

### *Methods*

Four studies were conducted that would each meet one or both of the objectives. Simulation was used to explore the number of patients and events required to limit the risk of overestimation of intervention effects to 'acceptable' levels. Empirical assessment was used to explore the performance of the popular measure of heterogeneity, $I^2$, and its associated 95% confidence intervals (CIs) as evidence accumulates. Empirical assessment was also used to compare inferential agreement between the widely used DerSimonian-Laird random-effects model and four alternative models. Lastly, a narrative review was undertaken to identify and appraise available methods for combining health related quality of life (HRQL) outcomes.

### *Results and conclusion*

The information required to limit the risk of overestimation of intervention effects is typically close to what is known as the optimal information size (OIS, i.e., the required meta-analysis sample size). $I^2$ estimates fluctuate considerably in meta-analyses with less than 15 trials and 500 events; their 95% confidence intervals provide desired coverage. The choice of random-effects

has ignorable impact on the inferences about the intervention effect, but not on inferences about the degree of heterogeneity. Many approaches are available for pooling HRQL outcomes. Recommendations are provided to enhance interpretability. Overall, each manuscript met at least one thesis objective.

## Preface

This thesis is a "sandwich thesis" consisting of four individual manuscripts as well as two additional analyses that further link the concepts and findings of these four manuscripts. At the time of writing (Oct 18 – 2011) two of the four individual manuscripts (chapters 2 and 5) have been published/accepted for publication in peer reviewed journals, and the remaining two (chapters 3 and 4) have been accepted for revision in peer reviewed journals. Kristian Thorlund's contributions to all the manuscripts and the additional analyses include: developing the research ideas and research questions, designing the studies, writing the protocol and analysis plan, performing the data extraction, conducting all statistical analyses, writing up all manuscripts, submitting the manuscripts; and responding to reviewers' comments. The work in this thesis was conducted between May 2009 and October 2011.

## Acknowledgements

I would first and foremost like to express my profound gratitude to my supervisors and thesis committee members, Dr. Lehana Thabane, Dr. PJ Devereaux, and Dr. Gordon Guyatt. You have each been tremendous sources of inspiration and absolutely fantastic role models. I am forever indebted to you for the education, opportunities and support you have given me.

A heartfelt thank you goes out to Dr. Christian Gluud, Dr. Jørn Wetterslev, and the rest of the *trial sequential analysis* team at Copenhagen Trial Unit. I truly treasure having worked with each of you and I hope you know how much your continuing support, friendship and collaboration have meant to me in the past and during my PhD. Christian and Jørn, I would not have been where I am today, were it not for you.

I like would to give a special thanks to Dr. Bradley Johnston and Dr. Michael Walsh for the great collaborations and discussions I have had with each of you during my time at McMaster. I would like to thank Shirley Petite and Andrea Robinson for always being extremely helpful and forthcoming. Also, thank you to all the co-authors of my thesis manuscripts for your great inputs and contributions; and thank you to all the administrative staff at HRM for doing your job so well. Last but not least, many thanks to all the great folks at CE&B and PHRI with whom I have shared offices, cubicle spaces or tables at the pub - life at Mac would just not have been the same without you.

# Table of Contents

## List of Tables

## List of Figures

# Chapter 1: Introduction

Meta-analysis and systematic reviews have become widely accepted and used in medical research over the past three decades.[1,2] Many methodological and practical challenges of meta-analysis and systematic reviews have been addressed in a substantial number of research studies and guidelines over the past years.[2-4] To a large extent, the increased acceptance of meta-analysis and systematic reviews as a valid scientific discipline may be attributed to these extensive methodological research efforts. Several tools and guidelines have been proposed and developed for bias risk assessment and reporting of results in systematic reviews, and many are now widely applied.[3-6] Somewhat surprising, however, the applied statistical methodology in the majority of published meta-analyses remains overly simple and out-dated, despite the fact that several improvements to the currently employed methods have been proposed and tested.[2] This seeming disproportionality between research efforts and advances in practice may in part be explained by the fact that statistics is extremely challenging to disseminate to the clinical audience. Considering that most systematic reviews, due to limited resources, do not have a statistician (or someone with adequate proficiency in statistics) on the author team, it seems crucial to ensure widespread dissemination of shortcomings associated with the currently employed (and out-dated) methods as well as the newer and relevant proposed statistical advances in meta-analysis.

**Problems with advancing statistical practice in meta-analysis**

Meta-epidemiological research, so far, has identified and demonstrated the importance of several determinants of the validity, reliability and interpretability of research findings in meta-analyses and systematic reviews. Most commonly, they have demonstrated associations between

overestimation of intervention effects and various sources of bias (methodological bias, publication bias, outcome reporting bias, etc.) and design features (use of composite endpoints, early stopping for benefit, etc.) among studies with inconsistent results.[7-14] Following the publication of such meta-epidemiological studies, experts have typically joined forces and developed tools and guidelines for assessing the quality of the evidence accordingly. Several statistical advances have also been made in meta-analysis, but most of these have rarely been applied in practice. The limited adoption of statistical advances in meta-analysis may likely be due to the sources through which they are being disseminated. That is, many statistical advances in meta-analysis have been published in advanced statistical and epidemiological journals like Statistics in Medicine and Journal of Clinical Epidemiology.[2] In such journals, manuscripts are often written in a highly technical language. While this is beneficial for researchers who want to get an in-depth understanding of the topic, it does little in educating the 'common' systematic review author. In addition, one could argue that many published meta-analysis statistical methodology papers have in fact devised methods for tackling data analysis problems which may not be relevant until 5, 10, or even 15 years from now (note: this thesis was submitted in July 2011). For example, proposing a hybrid of meta-regression and multiple treatment comparison (MTC) meta-analysis, although statistically novel, may not be relevant until the systematic review community has solved the issues of combining MTC data without MTC covariates.[15] Thus, it is not surprising that so many statistical advances in meta-analysis have not yet reached the clinical audience.

**Predictors of widespread dissemination of statistical advances**

Widespread dissemination and implementation of statistical advances in meta-analysis are just as important as widespread dissemination and implementation of other issues in meta-analysis. In brief, statistical issues in meta-analysis comprise the following: type I error and power, bias and precision (random error) associated with the estimation of all parameters and statistics being considered, and the appropriateness of the employed meta-analytic (statistical) model(s). Some examples of statistical advances in meta-analysis that have enjoyed widespread dissemination and implementation are the $I^2$ measure of the degree (percentage) of heterogeneity as well as indirect and multiple comparison methods. Both of these methods met a pressing demand in the meta-analysis and systematic review community. Each method provides relatively simple answers to complex problems that occur frequently in meta-analysis and their concepts are easily understood. Presumably, their successes can also partially be attributed to the degree to which they were disseminated. For the $I^2$ measure, the authors not only published a methods paper in Statistics in Medicine.[16] They also published a 'light' version in the British Medical Journal and described the method in the Cochrane Handbook.[4,17] Perhaps most importantly, the wide use of $I^2$ can be attributed to the fact that it was made an integral part of the Review Manager (RevMan) software used for all Cochrane systematic reviews and presumably for many systematic reviews published outside the Cochrane Library.[18] For indirect and multiple comparison methods, the methods seemed to catch on as 'the new hot thing' after a series of events, such as emphasis on these methods during recent Cochrane Collaboration colloquia, publication of multiple treatment comparisons on high impact topics, and a number of meta-epidemiological studies and educational papers.[19-34] Likewise, free user-friendly software for indirect comparisons has been released and several statistical code examples (e.g., WinBUGS code) have been published.[35] It

therefore seems that the recipe for wide dissemination of statistical advances in meta-analysis must include at least some of the elements in the above examples.

**Issues with validation of widely used statistical methods in meta-analysis**

While many statistical advances in meta-analysis do not enjoy widespread dissemination and implementation, ironically, those which do will typically not have undergone sufficient testing. The $I^2$ measure, for example, was implemented in RevMan at a time where only a few uncomprehensive simulations had suggested the measure had acceptable statistical properties.[16,36] More recently, some evidence has emerged to cast light over interpretational challenges associated with use of the $I^2$ measure.[37-41]

**Thesis objective**

There is a need for statistical contributions to meta-analysis where:

1) Improvements to current statistical practice in meta-analysis are conveyed at the level that most systematic review authors will be able to understand; and where:

2) Current statistical methods that are widely applied in meta-analytic practice undergo thorough testing and examination.

Of course, the research efforts needed to fill this demand would be substantial. The objective of this PhD thesis is to address some of this demand.

**Overview of approaches to achieving the objective**

This PhD thesis includes four separate research papers. Each paper deals with properties of currently applied methods for which their presumed merits have not been properly confirmed, or for which their statistical shortcomings have not been widely disseminated to the systematic review community. In case of the latter, alternatives and natural additions to the investigated methods are also studied and their performances are contrasted with the performance of the currently employed methods. For the design of each of the studies included in the PhD thesis, considerable attention was given to the need for simple performance measures which a clinical audience with basic biostatistics proficiency would be able to understand. This was done to meet objective 1, that is, to bridge some of the gaps between statistics in meta-analysis and the general systematic review audience. All papers deal with challenges in 'pair wise' meta-analysis, i.e., conventional meta-analysis comparing two interventions.

**Summary of chapters**

Chapter 2 deals with the risk of overestimation of intervention effects in binary meta-analysis due to imprecision (random error). It has previously been demonstrated empirically that imprecise meta-analyses (i.e. meta-analyses including a limited number of patients and events) are prone to yield large intervention effects estimates, and that such estimates tend to dissipate as more evidence is accumulated.[42-45] Such 'early' overestimates can particularly be compelling when backed up by a p-value smaller than 0.05. The GRADE working group and other authors have encouraged the use of the optimal information size (OIS, i.e. the required meta-analysis sample size) to assess the reliability of intervention effect estimates.[6,44,46-53] The OIS is being used increasingly in meta-analysis. Although the immediate statistical purpose of the OIS is to

provide a yardstick for the level of evidence which is required to achieve the desired type I error and type II error (or power) to detect some realistic intervention effect, authors of methodological papers and members of the GRADE working group have suggested that surpassing the OIS also provides protection against overestimation due to random error.[6,53] Theoretically this claim seems plausible, but nevertheless still needs validation. The study presented in chapter 2 is a comprehensive simulation study that explores the magnitude and likelihood of overestimation due to random error in relation to the cumulative number of patients and events in a meta-analysis. The number of patients and events required to limit the risk of overestimation to 'acceptable' levels are derived for the explored meta-analysis scenarios, and these estimates are contrasted with matching OIS calculations. The study presented in chapter 2 therefore serves to inform crucial unexplored issues of imprecision and use of OIS in meta-analysis before these are disseminated more widely.

Chapter 3 deals with the validity and reliability of the $I^2$ measure as evidence accumulates. Since $I^2$ became the most commonly used measure of heterogeneity in meta-analysis, a number of shortcomings and interpretational challenges associated with this measure have been identified.[37-41] Theoretically, all of these shortcomings may impact the validity and reliability of $I^2$ – especially when the evidence is limited. Much like the reliability of intervention effect estimates should be assessed in relation to some yardstick for adequate precision (like the OIS), $I^2$ should likewise be interpreted according to the level of evidence and the degree of uncertainty which surrounds it. The study presented in chapter 3 reviews the various factors that may impact the reliability and validity of $I^2$ as evidence accumulates (and particularly when evidence is sparse). The explanation of these factors caters to an audience with only basic knowledge of

statistical concepts. The performance of $I^2$ is explored empirically in 16 binary outcome meta-analyses with 'sufficient' evidence. This is done by plotting the cumulative $I^2$ estimates in relation to the cumulative number of events and trials. Because it is hypothesized that $I^2$ will incur substantial fluctuations over time, the performance of the 95% confidence intervals for $I^2$ over time is explored similarly. More specifically, it is assessed whether the 95% confidence intervals illustrate the appropriate degree of uncertainty associated with estimating $I^2$ and protect against spurious inferences about the degree of heterogeneity in meta-analysis. The study presented in chapter 3 therefore serves to cast further light on crucial shortcomings of the $I^2$ measure and promote simple aids for more accurate and reliable interpretation of this measure.

Chapter 4 challenges the widespread use of the conventional DerSimonian-Laird random-effects meta-analysis model when the included studies are heterogeneous. The only difference between the DerSimonian-Laird random-effects meta-analysis model and other 'conventional setup' random-effects meta-analysis models is the estimator which is used to estimate the between-study variance. The DerSimonian-Laird model uses, not surprisingly, the DerSimonian-Laird estimator.[54] Other models make use of other estimators.[55,56] Many comprehensive simulations studies have demonstrated that the DerSimonian-Laird estimator tends to underestimate the between-study variance.[55-62] When this occurs, the 95% confidence interval for the intervention effect estimate tends to become artificially narrow and heterogeneity estimates (e.g., $I^2$) become artificially small. Simulation studies have also demonstrated that alternative estimators, in many cases, perform better than the DerSimonian-Laird estimator.[55-62] Yet, these issues do not seem to have reached the systematic review community. The study presented in chapter 4 aims to test whether these issues are in fact problematic in practice. This is attained by comparing agreement

between the DerSimonian-Laird model and selected alternatives on measures of inference which all systematic review authors employ: the p-value for the intervention effect null hypothesis, 95% confidence interval for the estimated intervention effect, and the percentage of heterogeneity (in this study, measures which are conceptually similar to $I^2$ but specifically derived from the random-effects model in question). Agreement is assessed empirically for 920 'primary outcome' Cochrane meta-analyses.

Chapter 5 is the only one that deals with meta-analysis of continuous outcomes. Health related quality of life (HRQL) outcomes and patient reported outcomes (PROs) are becoming increasingly popular in clinical trials. Therefore, proper methods for combining such outcomes in meta-analysis are necessary.[63-66] Similar HRQL and PRO constructs are typically measured with different instruments across trials, and so combining them in a meta-analysis necessitates some transformation to a uniform scale. Typically the standardized mean difference, in which mean differences are reported in standard deviation units, has been used.[4] However, this method relies on standard deviations being comparable across trials (which is most often not the case). Further, in many areas of medicine, clinicians may find it difficult to interpret intervention effects in standard deviation units. The study presented in chapter 5 is a tutorial and review of 12 identified alternatives to the standardized mean difference. This review outlines the strengths and limitations associated with the identified methods, and explores their performance in two illustrative examples. Perhaps more importantly, the review provides simple recommendations based on a 2-step algorithm, about which method should be the preferred primary and complementary method for pooling and presenting continuous outcomes meta-analysis.

Chapter 6 presents some selected additional analyses that explore further links between chapters 2, 3 and 4. First, I explore how the degree of heterogeneity evolves over time with the alternative random-effects models (between-trial variance estimators) considered in chapter 4. Second, I explore whether some of the inferential discrepancies between random-effects models, which are observed in chapter 4, are more likely to occur in meta-analyses with sparse evidence compared to meta-analysis close to or beyond their OIS.

In chapter 7 (the discussion) the findings of chapters 2 to 6 are summarized. I discuss how well the manuscripts presented in chapters 2 to 5 meet the objectives of this thesis. I further discuss how well each of the manuscripts, when published, will aid in disseminating statistical methods currently not used (widely) in practice; as well as exploring the performance of widely used measures.

## References


(1) Lee WL, Baussel RB. The growth of health-related meta-analyses published from 1980 to 2000. *Evaluation and the Health Professions* 2001; 24:327-335.

(2) Sutton AJ, Higgins JPT. Recent developments in meta-analysis. *Statistics in Medicine* 2008; 27:625-650.

(3) Guyatt GH, Rennie D, Meade MO, Cook DJ. Users' guide to the medical literature: a manual for evidence-based clinical practice. 2nd ed. New York, NY: McGraw-Hill; 2008.

(4) Higgins JPT, Green S. Cochrane Handbook for systematic reviews of interventions, version 5.0.0. John Wiley & Sons; 2009.

(5) The EQUATOR network - Library of health research reporting. 2009. http://www.equator-network.org/resource-centre/library-of-health-research-reporting/

(6) Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008; 336(7650):924-926.

(7) Abraha I, Montedori A. Modified intention to treat reporting in randomised controlled trials: systematic review. *British Medical Journal* 2010; 340:c2697.

(8) Bassler D, Briel M, Montori VM, Lane M, Glasziou P, Zhou Q et al. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. *JAMA* 2010; 303(12):1180-1187.

(9) Chan AW, Hrobjartsson A, Haahr M, Gotzsche P, Altman D. Empirical evidence for selective reporting of outcomes in randomized trials. comparison of protocols to published articles. *Journal of American Medical Association* 2004; 291:2457-2465.

(10) Chan AW, Altman D. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *Canadian Medical Association Journal* 2004; 171:735-740.

(11) Dwan K, Alman D, Arnaiz J, Bloom J, Chan AW, Cronin E et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS Medicine* 2008; 3:e3081.

(12) Ferreira-Gonzalez I, Busse JW, Heels-Ansdell D, Montori VM, Akl EA, Bryant DM et al. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. *BMJ* 2007; 334(7597):786.

(13) 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 of Systematic Reviews* 2009;(1).

(14) Wood L, Egger M, Gluud LL, Schulz K, Jüni P, Gluud C et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemioogical study. *British Medical Journal* 2008; 336:601-605.

(15) Cooper NJ, Sutton AJ, Morris D, Ades AE, Welton NJ. Addressing between-study heterogeneity and inconsistency in mixed-treatment comparisons: Application to stroke prevention treatments in individuals with non-rheumatic atrial fibrilation. *Statistics in Medicine* 2009; 28(14):1861-1881.

(16) Higgins JPT, Thompson S. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 2002; 21:1539-1558.

(17) Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analysis. *British Medical Journal* 2003; 327(7414):557-560.

(18) Review Manager (RevMan) [Computer program]. Version 5.0. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2008.

(19) Ades AE. ISPOR States Its Position on Network Meta-Analysis. *Value Health* 2011; 14(4):414-416.

(20) Jansen JP, Fleurence R, Devine B, Itzler R, Barrett A, Hawkins N et al. Interpreting Indirect Treatment Comparisons and Network Meta-Analysis for Health-Care Decision Making: Report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: Part 1. *Value Health* 2011; 14(4):417-428.

(21) Hoaglin DC, Hawkins N, Jansen JP, Scott DA, Itzler R, Cappelleri JC et al. Conducting Indirect-Treatment-Comparison and Network-Meta-Analysis Studies: Report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices-Part 2. *Value Health* 2011; 14(4):429-437.

(22) Mills EJ, Druyts E, Ghement I, Puhan MA. Pharmacotherapies for chronic obstructive pulmonary disease: a multiple treatment comparison meta-analysis. *Clin Epidemiol* 2011; 3:107-129.

(23) Elliott WJ, Basu S, Meyer PM. Network meta-analysis of heart failure prevention by antihypertensive drugs. *Arch Intern Med* 2011; 171(5):472-473.

(24) Bangalore S, Kumar S, Kjeldsen SE, Makani H, Grossman E, Wetterslev J et al. Antihypertensive drugs and risk of cancer: network meta-analyses and trial sequential analyses of 324,168 participants from randomised trials. *Lancet Oncol* 2011; 12(1):65-82.

(25) Mills EJ, Wu P, Chong G, Ghement I, Singh S, Akl EA et al. Efficacy and safety of statin treatment for cardiovascular disease: a network meta-analysis of 170,255 patients from 76 randomized trials. *QJM* 2011; 104(2):109-124.

(26) Trikalinos TA, Alsheikh-Ali AA, Tatsioni A, Nallamothu BK, Kent DM. Percutaneous coronary interventions for non-acute coronary artery disease: a quantitative 20-year synopsis and a network meta-analysis. *Lancet* 2009; 373(9667):911-918.

(27) Strassmann R, Bausch B, Spaar A, Kleijnen J, Braendli O, Puhan MA. Smoking cessation interventions in COPD: a network meta-analysis of randomised trials. *Eur Respir J* 2009; 34(3):634-640.

(28) Mills EJ, Rachlis B, Wu P, Devereaux PJ, Arora P, Perri D. Primary prevention of cardiovascular mortality and events with statin treatments: a network meta-analysis involving more than 65,000 patients. *J Am Coll Cardiol* 2008; 52(22):1769-1781.

(29) Stettler C, Allemann S, Wandel S, Kastrati A, Morice MC, Schomig A et al. Drug eluting and bare metal stents in people with and without diabetes: collaborative network meta-analysis. *BMJ* 2008; 337:a1331.

(30) Jansen JP, Crawford B, Bergman G, Stam W. Bayesian meta-analysis of multiple treatment comparisons: an introduction to mixed treatment comparisons. *Value Health* 2008; 11(5):956-964.

(31) Song F, Loke YK, Walsh T, Glenny AM, Eastwood AJ, Altman DG. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. *BMJ* 2009; 338:b1147.

(32) Salanti G, Higgins JP, Ades AE, Ioannidis JP. Evaluation of networks of randomized trials. *Stat Methods Med Res* 2008; 17(3):279-301.

(33) O'Regan C, Ghement I, Eyawo O, Guyatt GH, Mills EJ. Incorporating multiple interventions in meta-analysis: an evaluation of the mixed treatment comparison with the adjusted indirect comparison. *Trials* 2009; 10:86.

(34) Song F, Harvey I, Lilford R. Adjusted indirect comparison may be less biased than direct comparison for evaluating new pharmaceutical interventions. *J Clin Epidemiol* 2008; 61(5):455-463.

(35) Indirect Treatment Comparison [computer program]. Version 1.0. Ottawa: Canadian Agency for Drugs and Technologies in Health [ 2009.

(36) Takkouche B, Cardarso-Suares C, Spiegelman D. Evaluation of old and new tests of heterogeneity in epidemiological meta-analysis. *American Journal of Epidemiology* 1999; 150(2):206-215.

(37) Ioaniddis JP. Interpretation of tests of heterogeneity and bias in meta-analysis. *Journal of Evaluation of Clinical Practice* 2008; 14:951-957.

(38) Ioannidis JP, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analysis. *BMJ* 2007; 335:914-916.

(39) Jackson D. The implications of publication bias for meta-analysis' other parameter. *Statistics in Medicine* 2006; 25(17):2911-2921.

(40) Rücker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I(2) in assessing heterogeneity may mislead. *BMC Medical Research Methodolgy* 2008;(8):79.

(41) Wetterslev J, Thorlund K, Brok J, Gluud C. Estimating required information size by quantifying diversity in a random-effects meta-analysis. *BMC Medical Research Methodology* 2009; 9(86).

(42) Gehr B, Weiss C, Porzsolt F. The fading of reported effectiveness. A meta-analysis of randomised controlled trials. *BMC Medical Research Methodology* 2006; 6:25.

(43) Ioannidis J, Lau J. Evolution of treatment effects over time: empirical insight from recursive meta-analyses. *Proc Nat Acad Sci U S A* 2001; 98(3):831-836.

(44) Thorlund K, Devereaux PJ, Wetterslev J, Guyatt G, Ioannidis JP, Thabane L et al. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? *International Journal of Epidemiology* 2009; 38:276-286.

(45) Trikalinos TA, Churchill R, Ferri M, Leucht S, Tuunainen A, Wahlbeck K et al. Effect sizes in cumulative meta-analyses of mental health randomized trials evolved over time. *Journal of Clinical Epidemiology* 2004; 57(11):1124-1130.

(46) Brok J, Thorlund K, Gluud C, Wetterslev J. Trial sequential analysis reveals insufficient information size and potentially false positive results in many meta-analyses. *Journal of Clinical Epidemiology* 2008; 61:763-769.

(47) Brok J, Thorlund K, Wetterslev J, Gluud C. Apparently conclusive meta-analyses may be inconclusive - Trial sequential analysis adjustment of random error risk due to repetitive testing of accumulating data in apparently conclusive neonatal meta-analyses. *International Journal of Epidemiology* 2009; 38:287-298.

(48) Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. *Journal of Clinical Epidemiology* 2008; 61:64-75.

(49) Wetterslev J, Thorlund K, Brok J, Gluud C. Estimating required information size by quantifying diversity in random-effects model meta-analyses. *BMC Med Res Methodol* 2009; 9(1):86.

(50) Daya S. Optimal Information Size. *Evidence-based Obstetrics and Gynecology* 2002; 4:53-55.

(51) Pogue J, Yusuf S. Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. *Controlled Clinical Trials* 1997; 18:580-593.

(52) Pogue J, Yusuf S. Overcoming the limitations of current meta-analysis of randomised controlled trials. *Lancet* 1998; 351:47-52.

(53) Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D et al. GRADE guidelines: 6. rating the quality of evidence - precision. *Journal of Clinical Epidemiology* 2011; (In Press).

(54) DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; 7:177-188.

(55) Sánchez-Meca J, Marín-Martínez F. Confidence intervals for the overall effect size in random-effects meta-analysis. *Psych Meth* 2008; 1:31-38.

(56) Sidik K, Jonkman JN. A comparison of heterogeneity variance estimators in combining resuls of studies. *Stat Med* 2007; 26:101-129.

(57) Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. *Stat Med* 2000; 20:825-840.

(58) Brockwell SE, Gordon IR. A simple method for inference on an overall effect in meta-analysis. *Stat Med* 2007; 26:4531-4543.

(59) Hartung J, Makambi KH. Reducing the number of unjustified significant results in meta-analysis. *Communications in Statistics - Simulation and Computation* 2003; 32(4):1179-1190.

(60) Makambi KH. The effect of the heterogeneity variance estimator on some tests of efficacy. *J Biopharm Stat* 2004; 14(2):439-449.

(61) Sidik K, Jonkman JN. Simple heterogeneity variance estimation for meta-analysis. *J Roy Stat Soc (C)* 2005; 54:367-384.

(62) Sidik K, Jonkman J. Robust variance estimation for random-effects meta-analysis. *Comp Stat Data An* 2006; 50:3681-3701.

(63) Valderas JM, Kotzeva A, Espallargues M, Guyatt G, Ferrans CE, Halyard MY et al. The impact of measuring patient-reported outcomes in clinical practice: a systematic review of the literature. *Qual Life Res* 2008; 17(2):179-193.

(64) Angst F, Verra ML, Lehmann S, Aeschlimann A. Responsiveness of five condition-specific and generic outcome assessment instruments for chronic pain. *BMC Medical Research Methodology* 2008; 8(26).

(65) Wiebe S, Guyatt GH, Weawer B, Matijevic S, Sidwell C. Comparative responsiveness of generic and specific quality of life instruments. *Journal of Clinical Epidemiology* 2003; 56:52-60.

(66) Patient Reported Outcomes and Quality of Life Instruments Database. 2011. http://www.proqolid.org/proqolid/search__1/pathology_disease?pty=1924

# Chapter 2: The number of patients and events required to limit the risk of overestimation of intervention effects in meta-analysis – a simulation study

Authors:

Kristian Thorlund

Georgina Imberger

Michael Walsh

Rong Chu

Christian Gluud

Jørn Wetterslev

Gordon Guyatt

Philip J Devereaux

Lehana Thabane

Word count

Abstract: 318

Manuscript: 3725

## Abstract

**Background:**

Meta-analyses including a limited number of patients and events are prone to yield overestimated intervention effect estimates. While many assume bias is the cause of overestimation, theoretical considerations suggest that random error may be an equal or more frequent cause. The independent impact of random error on meta-analyzed intervention effects has not previously been explored. It has been suggested that surpassing the optimal information size (i.e., the required meta-analysis sample size) provides sufficient protection against overestimation due to random error, but this claim has not yet been validated.

**Methods:**

We simulated a comprehensive array of meta-analysis scenarios where no intervention effect existed (i.e., relative risk reduction (RRR) = 0%) or where a small but possibly unimportant effect existed (RRR = 10%). We constructed different scenarios by varying the control group risk, the degree of heterogeneity, and the distribution of trial sample sizes. For each scenario, we calculated the probability of observing overestimates of RRR > 20% and RRR > 30% for each cumulative 500 patients and 50 events. We calculated the cumulative number of patients and events required to reduce the probability of overestimation of intervention effect to 10%, 5%, and 1%. We calculated the optimal information size for each of the simulated scenarios and explored whether meta-analyses that surpassed their optimal information size had sufficient protection against overestimation of intervention effects due to random error.

**Results:**

The risk of overestimation of intervention effects was usually high when the number of patients and events was small and this risk decreased exponentially over time as the number of patients and events increased. The number of patients and events required to limit the risk of overestimation depended considerably on the underlying simulation settings. Surpassing the optimal information size generally provided sufficient protection against overestimation.

**Conclusions:**

Random errors are a frequent cause of overestimation of intervention effects in meta-analyses. Surpassing the optimal information size will provide sufficient protection against overestimation.

## Introduction

Systematic reviews and meta-analyses combining evidence from several high-quality randomized clinical trials (RCTs) are generally considered the highest level of evidence for effects of interventions.[1–3] Many systematic reviews address questions important and pressing to a large group of patients and clinicians. Therefore, these analyses are often conducted at a stage when the evidence on a particular question is still limited. Such meta-analyses lack the precision (i.e., are underpowered) to establish realistic intervention effects with a high level of confidence.[4–11] Yet, it is not infrequently that such preliminary meta-analyses yield apparently large intervention effect estimates which, if meeting the conventional criterion for statistical significance (i.e., $p \leq 0.05$), can appear compelling.[4–11] Empirical studies suggest that when more evidence is accumulated over time, many of these 'early' large apparent intervention effects turn out to be substantial overestimates.[4–6,12] Meta-analysis authors often assume that time-lag, publication bias, methodological bias, or outcome reporting bias are the main cause(s) of early overestimation, but theoretical considerations suggest that lack of precision may be an equally or more frequent cause.[3–11,13]

As authors and users of meta-analyses and systematic reviews, we wish to avoid the mistake of trusting spuriously large meta-analyzed intervention effects. Because precision (and power) is highly correlated with the cumulative number of patients and events, some authors have recommended that meta-analyzed intervention effect estimates should be interpreted in relation to the cumulative number of patients or events.[6–9,14–17] In particular, a *required* or an *optimal* information size (OIS, analogous to a required sample size in a clinical trial) has been proposed for meta-analysis.[9,15–17] While we find this proposal highly useful, the optimal information size does not provide insight into the degree and likelihood of overestimation of intervention effects

that one can expect at various preceding stages of a meta-analysis. Further, it is unknown whether conventional information size requirements (i.e., $\alpha = 5\%$, $\beta = 10\%$, and plausible a priori assumptions about the intervention effect, control group risk, and degree of heterogeneity), provide sufficient protection against overestimation of meta-analyzed intervention effects caused by random errors (imprecision). The existing empirical studies on this topic are, unfortunately, limited by their respective sample sizes (the number of meta-analyses studied empirically), and thus, do not provide a reliable basis for assessing the expected degree and likelihood of overestimation at various stages evidence accumulation. Further, because the impact of bias (systematic error) is next to impossible to infer with certainty in individual meta-analyses, it is also difficult to isolate the degree to which random error alone (and not bias) causes overestimation in individual meta-analyses. The sole effect of random error on the meta-analyzed intervention effect can, however, be accurately evaluated via simulation.

To assess the degree and likelihood with which imprecision causes overestimation of intervention effects at various stages of a meta-analysis, we undertook a simulation study. We measured the probability of observing relative risk reduction estimates that could potentially represent important overestimations after every 500 or 200 patients and for every 50 or 20 events (depending on the simulation scenario). We explored how well conventional information size requirements protected against overestimation by comparing these with the number of patients and events required for reducing the probability of overestimation to 'acceptable levels' (i.e., 10%, 5%, or 1%). Our simulations cover a comprehensive array of scenarios that approximate common meta-analysis data sets and our tables and figures may readily aid systematic review

authors in assessing the risk of overestimation due to random error in their specific meta-analysis.

## Methods

### Simulation framework

We simulated binary meta-analysis data sets using a DerSimonian-Laird random-effects model framework.[3,18,19] The statistical formulation for the random-effects model as well as the formula for the DerSimonian-Laird estimator for the between-trial variance are presented in the supporting information (Appendix S1). We simulated meta-analysis scenarios based on assumed distributions and fixed, chosen values for the trial specific variables: the trial sample sizes, the control group risks, the 'true' intervention effect, and the degree of heterogeneity. We used two trial sample size distributions: one based on a survey of the Cochrane Heart Group meta-analyses on mortality (see supporting information Table S1 and Table S2) and one based on our subjective assessment of what constitutes a 'common' meta-analysis scenario. We used four different uniform distributions to sample the control group risk: 1% to 5% (representing 'low' control group risk), 5% to 15% (representing 'moderately low'), 15% to 40% (representing 'moderate'), and 40% to 80% (representing 'high'). We used three different values of the between-trial variance (referred to as $\tau^2$ in the supporting information - Appendix S1) of the log relative risk to simulate different degrees of heterogeneity: 0.05, 0.15, and 0.25. Because our study objective was to investigate various aspects of overestimation of intervention effects, we used relative risk reduction (RRR) = 0% (no effect) and RRR = 10% (small but possibly unimportant effect) as the 'true' underlying intervention effects. In-depth rationale for the choice of the performed simulation scenarios is provided in appendix S2 in the supporting information.

Further, the technical details of our simulation approach are described in detail in Appendix S2 in the supporting information.

For each scenario, we simulated 20,000 meta-analysis data sets, and for each simulated meta-analysis data set, we simulated 100 trials. Although meta-analysis data sets including this many trials are uncommon in practice, we were interested in estimating the risk of overestimation both in common as well as uncommon meta-analysis scenarios. Simulating 100 trials for each meta-analysis data set allowed us to accurately estimate the risk of overestimation regardless of the cumulative number of patients and events. Figure 1 presents a flowchart of the simulation and analysis structure.

**The optimal information size**

The optimal information size, OIS, for a binary outcome meta-analysis (also referred to as the required information size) is calculated as

$$OIS = (4 \cdot (z_{1-\alpha} + z_{1-\beta})^2 \cdot P \cdot (1-P)/\delta^2) \cdot (1/(1-I^2))$$

Where $z_{1-\alpha}$ and $z_{1-\beta}$ are the $(1-\alpha)$th and $(1-\beta)$th percentiles from the standard normal distribution, P is the average of the control group risk, $P_C$, and intervention group risk, $P_E$, $\delta$ is the difference between $P_C$ and $P_E$, and $I^2$ is the popular (heterogeneity) measure for the proportion variation in a meta-analysis explained by differences between trials rather than sampling error. (Note, $I^2$ is typically reported as a percentage (e.g., $I^2 = 35\%$), but in the OIS formula above, $I^2$ is a proportion (e.g., $I^2 = 0.35$)). The OIS provides the required number of patients in a meta-analysis to ensure that the maximum type I error is no larger than $\alpha$ and the maximum type II error is no

larger than β when testing for statistical significance. The OIS can be converted to the required number of events by multiplying the required number of patients by P (assuming an approximately equal number of patients in the two groups).

## Analysis

For each simulation scenario of 20,000 cumulative meta-analyses data sets, we recorded the DerSimonian-Laird random-effects model cumulative meta-analyzed RRR (1 minus the meta-analyzed relative risk), the cumulative number of patients, and the cumulative number of events after each included trial. For each simulation set (i.e., true RRR = 0% and true RRR = 10%), we judged that RRR estimates larger than 20% and 30% could potentially represent important overestimates. At any given cumulative number of patients and events, we therefore calculated the proportion of simulated meta-analysis RRR that were larger than these thresholds.

We assessed the degree and likelihood of overestimation at various stages of a meta-analysis. For each scenario, we plotted the proportion of overestimates (according to the two thresholds) in relation to the cumulative number of patients and events. For each plot, we divided the cumulative number of patients into intervals of 500 or 200 (depending on the scenario), and the cumulative number of events into intervals of 50 or 20 (depending on the scenario).

We assessed how many patients and events were required to reduce the proportion of overestimates to acceptable levels, according to the two thresholds. We calculated the number of patients and events required to limit the probability of overestimation (according to the two

thresholds) by 10%, 5%, and 1% - each of which could potentially constitute an 'acceptable' risk of overestimation.

We assessed the extent to which conventional information size requirements protect against overestimation. We calculated the optimal information sizes based on $\alpha = 5\%$ and $\beta = 20\%$, 10%, or 5%, with assumed control group risks set to the averages of the four control group risk distributions used in the simulation (i.e., $P_C = 3.0\%$, $P_C = 10.0\%$, $P_C = 27.5\%$, or $P_C = 60.0\%$), powered to detect intervention effects of RRR = 30% or RRR = 20%, and with heterogeneity corrections of $I^2 = 0.00$, $I^2 = 0.25$, or $I^2 = 0.50$ (corresponding to $I^2 = 0\%$, $I^2 = 25\%$, and $I^2 = 50\%$). In total, 72 OIS estimates were calculated. We then compared the calculated information size requirements with the simulation results by matching OIS estimates with the scenarios where the underlying assumptions were similar. For example, the estimated probabilities of overestimation from the simulation based on a control group risk between 5% and 15% and $tau^2 = 0.15$ were compared to the information size requirements based on an assumption of a 10% control group risk and 25% heterogeneity ($I^2 = 0.25 = 25\%$). For the comparison of information size requirements and simulation results, we post hoc created three categories for the 'acceptability' of the risk of overestimation: 'good', 'very good', and 'excellent'. We defined 'good' acceptability as the situation where the probability of observing an RRR > 20% was smaller than 10% and the probability of observing an RRR > 30% was smaller than 5%. We defined 'very good' acceptability as the situation where the probability of observing an RRR > 20% was smaller than 5% and the probability of observing an RRR > 30% smaller than 1%. Lastly, we defined 'excellent' acceptability as the situation where the probability of observing an RRR > 20% was smaller than 1%.

Of note, we did not record the probability of underestimation (i.e., we took a one-sided approach). Thus, 50% is the maximum observable probability of overestimation of intervention effects, and our results should be interpreted accordingly.

## Results

In most scenarios, the probability of overestimation was higher than 25% when the number of patients (or events) was small, but subsequently decreased exponentially (the x-axis is log scaled in figure 2, and in figures S1 to S12 in the supporting information).

Figure 2 presents the probability of overestimation in relation to the cumulative number of patients and events for a selected simulation scenario: no true intervention effect (RRR = 0%), moderate control group event risk (uniform distribution from 5% to 15%), and moderate heterogeneity (between-trial variance $\tau^2 = 0.15$), and distribution of trials sizes based on our survey of the Cochrane Heart Group meta-analyses. Figures S1 to S12 in the supporting information present the probability of overestimation in relation to the cumulative number of patients and events for all simulation scenarios.

The number of patients and events required for the probability of overestimation to drop below 10%, 5%, and 1% in the simulated scenarios are presented in Table 1, and Tables S3 and S4 in the supporting information. Table 1 presents the scenarios where the distribution of trial sample sizes were based on our survey of the Cochrane Heart Group meta-analyses, and Tables S3 and S4 in the supporting information present the scenarios where the distribution of trial sample sizes

was based on our assessment of what we subjectively assessed constituted a 'common' meta-analysis scenario.

The number of patients and events required to limit the risk of overestimation depended on the threshold for overestimation (i.e., RRR = 20% or RRR = 30%) and all the considered simulation components: relative risk reduction, control group risk, heterogeneity, and trial size distribution. The larger the overestimation (i.e., the larger the difference between the meta-analyzed and the true RRR), the smaller the number of patients and events required to limit the risk of overestimation. A larger number of patients was required to limit the risk of overestimation in the scenarios where the control group risk was low. Conversely, a smaller number of events was required to limit the risk of overestimation in the scenarios where the control group risk was low. The number of patients and events required to limit the risk of overestimation was generally smaller in scenarios when heterogeneity was set at the lowest level ($\tau^2 = 0.05$) than when it was set to the highest level ($\tau^2 = 0.25$). In contrast, in scenarios with the 'common' trial size distribution and with low control group risks (1-5%), the number of patients and events required was higher when heterogeneity was lowest. This reversed pattern was also observed in a few other isolated scenarios.

Table 2 presents the calculated optimal information size for 72 different settings (see analysis section for more detail). Table 3 and 4 present the number of patients and events required to limit the risk of overestimation, grouped by control group risk and distribution of trial sample size. The calculated OIS are included in these tables for comparison. In scenarios with low control group risk (1%-5%), the risk of overestimation generally reached very good or excellent

acceptability before reaching optimal information sizes (based on 80% power or 90% power). In scenarios with moderately low control group risk (5% to 15%), good acceptability was commonly reached before or close to the OIS based on 80% power, whereas very good and sometimes excellent acceptability was reached before the OIS based on 90% power or 95% power. In scenarios with moderate control group risk (15% to 40%), good acceptability was reached before the OIS based on 80% power and very good acceptability was usually reached before the OIS based on 95% power. In scenarios with high control group risk (40% to 80%), good acceptable was often (but not always) reached before the OIS based on 95% power. Some exceptions were observed in all of the above generalizations when the heterogeneity was large (i.e., $\tau^2 = 0.25$).

## Discussion

Our simulations provide valuable insight on the risk of overestimation of intervention effects in meta-analysis due to random errors over time. The risk of observing overestimated intervention effects due to random error at 'early' stages of a meta-analysis is substantial. The number of patients and events required to limit this risk depend considerably on each of the components considered in our simulation study: the degree of overestimation that is considered to be important, the underlying true effect, the control group risk, the degree of heterogeneity, and the distribution of trial sample sizes. However, the comparison of our simulation results with the approximately corresponding information size requirements demonstrated that upon reaching the OIS in a meta-analysis, one can be relatively confident that the intervention effect is not overestimated due to random error.

Our study comes with several strengths and limitations. Our simulations covered a wide spectrum of meta-analysis scenarios which we believe occur frequently in the systematic review literature. Our simulation results therefore have good generalizability to meta-analysis in practice. While the spectrum of scenarios covered in our simulations is not as extensive as seen in some previous meta-analysis simulation studies, adding additional scenarios to the current study would likely increase the complexity and hamper the interpretability of our findings. We believe the chosen spectrum of our simulations constitute a close to optimal balance between interpretability and generalizability.

Our simulation study is the first of its kind to contrast the risk of overestimation of intervention effects due to random errors with information size requirements. The statistical purpose of calculating the OIS is to gain control over the risk of obtaining a false positive finding (type I error) and a false negative finding (type II error). Extending this purpose, authors have previously considered information size requirements as a means of gaining control over the risk of overestimation.[2,20] Our simulation study is the first to explore the validity of this theoretical claim. However, we only investigated the extent to which information size requirements protect against overestimation when the underlying assumptions (e.g., a priori assumed RRR and control group risk) matched the parameter settings in a given simulation scenario (e.g., the assumed control group risk for the optimal information size was set to 10% when the control group risk in the simulation was sampled from a uniform distribution between 5% and 15%). That is, our findings hold for information sizes that have been calculated using the appropriate assumptions for a given scenario. In reality, it can be difficult to know which assumptions are most appropriate when doing information size calculations for a meta-analysis. The implications of

employing overly lenient or conservative a priori assumptions for the OIS are, theoretically, relatively straightforward. Lenient assumptions (e.g., $\beta = 20\%$ and RRR = 0.35) will result in relatively small information size requirements, and thus, an inappropriately high degree of confidence that the estimated intervention effect can be trusted (i.e., is not an overestimate). Conversely, conservative assumptions (e.g., $\alpha = 0.1\%$ and RRR = 0.10) have the potential to remove confidence about an intervention effect estimate, even if the intervention effect estimate is in fact reliable.

We mentioned in the introduction that various types of bias (e.g., methodological bias or publication bias) may also be important causes of overestimation of intervention effects.[3,13] We did not attempt to include any biases in our simulations. It is likely that when bias is present in a meta-analysis, a larger number of patients and events will be required to limit the risk of overestimation. In some cases, bias may limit the reliability of the size of the intervention estimate independent of how large the meta-analysis is.

Another limitation of our simulations is that the underlying true trial effects were sampled as random effects. This approach does not consider the possibility that the magnitude of trial effects to some extent may depend on time. For example, the first series of trials in a meta-analysis compared to the later trials may generally recruit a broader or narrower population, use shorter or longer follow-up, or administer higher or lower doses of a drug. Depending on the effect such time dependencies have on the evolution of the meta-analyzed intervention effect, the number of patients and events required to limit overestimation may be either larger or smaller than our results indicate.

Our simulation results are consistent with the results of previous empirical studies. More specifically, the pooled intervention effect estimates tend to fluctuate considerably when the number of patients and events are sparse, thus creating a high risk of overestimation.[4–6,12] Ioannidis and Lau previously investigated convergence of intervention effects in two fields, interventions in pregnancy and perinatal medicine and management of myocardial infarction. They found that more than 10,000 patients were generally required to relieve uncertainty about subsequent changes in meta-analyzed intervention effects.[4] Trikalinos et al. performed a similar study on interventions within the field of mental health and found that only 2000 patients were required to relieve uncertainty about subsequent changes in meta-analyzed intervention effects.[5] The meta-analyses considered by Ioannidis and Lau were similar to many of our simulated scenarios where the control group risk was 'low' and 'moderately low'. The meta-analyses considered by Trikalinos et al. were similar to many of our simulated scenarios where the control group risk was 'moderate' or 'high'.

The results of our simulation study have several implications. First, they underscore the need for information size requirements in all meta-analyses. Second, they illustrate the dangers of relying on intervention effect estimates before the OIS is reached (or is close to being reached), even when the presence of bias is unlikely. The figures in the supporting information provide meta-analysts with an opportunity to check the approximate risk of overestimation due to random error in their meta-analyses.

Two key inferential measures in a meta-analysis are the p-value and the 95% confidence interval associated with the estimated intervention effect. We wish to offer additional caution in interpreting meta-analyzed intervention effect estimates in the face of limited evidence. Large effect estimates (true or false) do not require high precision to reach conventional statistical significance (i.e., $p \leq 0.05$). As demonstrated in empirical studies, early large intervention effects are likely to dissipate and early statistically significant meta-analyses are likely to be false positives.[4–6,12,21] Therefore, when observing a large statistically significant intervention effect estimate (e.g., RRR > 30%) in a meta-analysis including a limited number of patients and events, one should always consider whether the meta-analysis, with the same precision, would have been statistically significant had the observed intervention effect been moderate or small. Chances are it would not. By the same token, one should always consider what values the confidence interval would have included had the effect estimate been moderate or small.

Even if an 'early' large intervention effect estimate is not supported by formal statistical significance, the situation may still be problematic. Large intervention effects will encourage clinical trial investigators to conduct further trials, and systematic review authors to perform regular updates of the meta-analysis until it either reaches statistical significance or the early trend has been definitively refuted. Updates of meta-analysis cause multiplicity due to repeated significance testing – a conduct which has been documented as highly problematic.[6,10,22–24] In particular, multiple testing increases the risk of observing a falsely significant result before the optimal information size has been surpassed. This may very well happen at a point where the risk of overestimation is still substantial. Moreover, in the face of repeated significance testing, confidence intervals suffer from reduced coverage, and thus an increased risk of precluding the

'true' intervention effect. Multiplicity due to repeated significance testing in meta-analysis can be accounted for by employing sequential testing procedure like the O'Brien-Fleming group sequential boundaries (i.e., adjusted thresholds for statistical significance) and adjusted confidence intervals can be constructed accordingly. Evidence suggests that these techniques provide adequate protection against false positives.[6,8,14,22] Given that such adjusted significance thresholds and the corresponding adjusted confidence intervals are tied to the calculated information size requirement, and given that information size criteria seem to provide adequate protection against 'early' overestimation, it seems reasonable to believe that adjusted significance thresholds and confidence intervals are appropriate inferential measures for interpreting early intervention effect estimates in meta-analysis.

In conclusion, the risk of overestimated intervention effects in meta-analysis due to random error is often substantial in the face of a limited number of patients and events. Insisting that a meta-analysis meets a reasonable OIS will ensure an acceptably low risk of observing an overestimated intervention effect due to random errors.

## References


(1) Guyatt G, Haynes RB, Jaeschke RZ, Cook DJ, Green L, et al. Users' Guides to the Medical Literature: XXV. Evidence-based medicine: principles for applying the Users' Guides to patient care. *JAMA* 2002; 284: 1290-1296.

(2) Guyatt G, Oxman AD, Vist GE, Kunz R., Falck-Ytter Y, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008; 336: 924-926.

(3) Higgins JP and Green S Cochrane Handbook for systematic reviews of interventions, version 5.0.0. John Wiley & Sons; 2009.

(4) Ioannidis J, Lau J Evolution of treatment effects over time: empirical insight from recursive meta-analyses. *Proc Nat Acad Sci USA* 2001; 98: 831-836.

(5) Trikalinos TA, Churchill R, Ferri M, Leucht S, Tuunainen A, et al. Effect sizes in cumulative meta-analyses of mental health randomized trials evolved over time. *Journal of Clinical Epidemiology* 2004; 57: 1124-1130.

(6) Thorlund K, Devereaux PJ, Wetterslev J, Guyatt G, Ioannidis JP, et al. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? *International Journal of Epidemiology* 2009; 38: 276-286.

(7) Brok J, Thorlund K, Gluud C, Wetterslev J. Trial sequential analysis reveals insufficient information size and potentially false positive results in many meta-analyses. *Journal of Clinical Epidemiology* 2008; 61: 763-769.

(8) Brok J, Thorlund K, Wetterslev J, Gluud C. Apparently conclusive meta-analyses may be inconclusive - Trial sequential analysis adjustment of random error risk due to repetitive testing of accumulating data in apparently conclusive neonatal meta-analyses. *International Journal of Epidemiology* 2009; 38: 287-298.

(9) Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. *Journal of Clinical Epidemiology* 2008; 61: 64-75.

(10) Berkey C, Mosteller F, Lau J. Uncertainty of the time of first sifnificance in random-effects cumulative meta-analysis. *Controlled Clinical Trials* 1996; 17: 357-371.

(11) Borm GF, Donders AR. Updating meta-analyses leads to larger type I error than publication bias. *Journal of Clinical Epidemiology* 2009; 62: 825-830.

(12) Gehr B, Weiss C, Porzsolt F. The fading of reported effectiveness. A meta-analysis of randomised controlled trials. *BMC Medical Research Methodology* 2006; 6: 25.

(13) Wood L, Egger M, Gluud LL, Schulz K, Jüni P, et al. (2008) Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemioogical study. British Medical Journal 336: 601-605.

(14) Devereaux PJ., Beattie WS, Choi PT., Badner NH, Guyatt GH, et al. How strong is the evidence for the use of perioperative beta blockers in non-cardiac surgery? Systematic review and meta-analysis of randomised controlled trials. *British Medical Journal* 2005; 331: 313-321.

(15) Pogue J, Yusuf S. Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. *Controlled Clinical Trials* 1997; 18: 580-593.

(16) Pogue J, Yusuf S. Overcoming the limitations of current meta-analysis of randomised controlled trials. *Lancet* 1998; 351: 47-52.

(17) Wetterslev J, Thorlund K, Brok J, Gluud C. Estimating required information size by quantifying diversity in a random-effects meta-analysis. *BMC Medical Research Methodology* 2009; 9.

(18) DerSimonian L, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; 7: 177- 188.

(19) Sidik K, Jonkman J. Simple heterogeneity variance estimation for meta-analysis. *Journal of Royal Statistical Society(C)* 2005; 54: 367-384.

(20) Guyatt G, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE Guidelines: 6. Rating the quality of evidence – imprecision. *Journal of Clinical Epidemiology* 2011; [Epub ahead of print].

(21) Pereira TV, Ioannidis J. Statistically significant meta-analyses of clinical trials have modest credibility and inflated effects. *Journal of Clinical Epidemiology* 2001; doi: 10.1016/j.jclinepi.2010.12.012.

(22) Higgins JP, Whitehead A, Simmonds M. Sequential methods for random-effects meta-analysis. *Statistics in Medicine* 2011; 30: 903-921.

(23) Lau J, Schmid C, Chalmers T. Cumulative meta-analysis of clinical trials builds evidence for exemplary medical care. *Journal of Clinical Epidemiology* 1995; 48: 45-57.

(24) van der Tweel I, Bollen C. Sequential meta-analysis: an efficient decision making tool. *Clinical Trials* 2010; 7: 136-146.

(25) Glasziou PP, Sheppard S, Brassey J. Can we really rely on the best trial? A comparison of individual trials and systematic reviews. *BMC Medical Research Methodology* 2010; 10.

(26) Nuesch E, Trelle S, Reichenbach S, Rutjes AWS, Tschannen B, et al. Small study effects in meta-analysis of osteoarthritis trials: meta-epidemiological study. *British Medical Journal* 2011; 341: c3515.

(27) Chan AW, Altman DG. Epidemiology and reporting of randomized clinical trials published in PubMed journals. *Lancet* 2005; 365: 1159-1162.

(28) Gluud C. The culture of designing hepato-biliary randomised clincial trials. *Journal of Hepatology* 2006; 44: 607-615.

(29) Brockwell S, Gordon I. A comparison of statitical methods for meta-analysis. *Statistics in Medicine* 2001 20: 825-840.

(30) Sidik K, Jonkman J. A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in Medicine* 2007; 26: 1964-1981.

## Supporting information 1 (Appendix S1)

### Random-effects model meta-analysis

In the meta-analytic framework, the random-effects model is defined as follows.[3,18,19] Assume we have k independent trials. Let $Y_i$ be the estimate of the effect from the individual trials. Let $\mu_i$ be the true intervention effect of the *i*th trial, and let $\sigma_i^2$ denote the variance of $\mu_i$. The trial specific intervention effects are assumed to vary across trials, with an underlying true effect, $\mu$, and a between-trial variance $\tau^2$. The random-effects model is defined hierarchically by

$$Y_i = \mu_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma_i^2)$$

$$\mu_i = \mu + E_i, \qquad E_i \sim N(0, \tau^2)$$

Collapsing the hierarchy, the observed effect measure, $Y_i$, is then assumed to satisfy the distributional relationship $Y_i \sim N(\mu, \sigma_i^2 + \tau^2)$, and the trial weights, $w_i^*$, are set as the inverse of the individual trial variances, $w_i^* = (\sigma_i^2 + \tau^2)^{-1}$. Here we use the asterix to indicate that the weights come from the random-effects model (as opposed to the weights coming from a fixed-effect model $w_i = \sigma_i^{-2}$). In practice, neither $\sigma_1^2$, …, $\sigma_k^2$, nor $\tau^2$ are known. The within-trial variances are typically often estimated using the sampling variances and the between-trial variance $\tau^2$ is typically obtained using some estimator (see below).

In the random-effects model meta-analysis the overall intervention effect is obtained as a weighted average of the observed intervention effects in the included trials

$$\mu_w = (\Sigma_i \, w_i^* \cdot Y_i)/(\Sigma_i \, w_i^*)$$

and the variance is estimated as

$$\mathrm{Var}(\mu_w) = 1/(\Sigma_i \, w_i^*)$$

***DerSimonian-Laird random-effects meta-analysis***

In the conventional random-effects model approach proposed by DerSimonian and Laird (DL), the between-trial variance is estimated using a method of moments based estimator.[18] Cochran's homogeneity test statistic, $Q = \Sigma \, w_i \, (Y_i - \mu_w)^2$, is used as the basis of the DL estimator, as its 1st moment takes the form $E(Q) = (k-1) + \tau^2 \, (S_1 - (S_2 / S_1))$, where $S_r = \Sigma \, w_i^r$, for r = 1,2. Isolating for $\tau^2$ then yields the expression for the method of moments estimator of the between-trial variance

$$\tau_{DL}^2 = \max(0, (Q - k + 1) / (S_1 - (S_2 / S_1)))$$

## Supporting information S2 (Appendix S2)

## Simulation

We aimed to create 'realistic' meta-analysis data sets. We first did so by confining our attention to one area of medicine, cardiology, and surveyed all meta-analyses on mortality from the Cochrane Heart Group (Issue 4, 2009) to inform 'realistic' simulation parameter settings. However, subsequently realizing an important lack of generalizability from this approach (as a peer reviewer kindly pointed out), we added a number of simulation parameter values and settings to allow for inferences that apply to a greater spectrum of meta-analysis scenarios.

When surveying the meta-analyses from the Cochrane Heart Group, we selected reviews that meta-analyzed results on mortality, including at least 3 trials and 100 events. We separately surveyed meta-analyses in which the median follow-up across trials was between 1 month and 1 year, and between 1 year and 5 years. We did not include meta-analyses including cluster randomized trials in our survey. For each eligible meta-analysis, we recorded the trial sample sizes, the trial control group risk, the pooled relative risks, the associated 95% confidence intervals, and the estimated DerSimonian-Laird between-trial variance (on a log relative risk scale, see Appendix S1 in supporting information). For each eligible meta-analysis, we also recorded the median, minimum, maximum, and the quartiles for the trial sample sizes and control group risks, and summarized these statistics in a table. With the surveyed information, we plotted all trial sample sizes as a histogram.

Tables S1 and S2 and Figure S13 in the supporting information present the data summaries produced from the survey of the Cochrane Heart Group meta-analyses that were used to inform the settings of our simulations.

**Determining simulation parameters**

We simulated binary data meta-analysis scenarios based on distributional assumptions for the trial specific variables: the underlying true intervention effect, $\mu$, the distribution of trial sample sizes, $n_i$, the observed control group risks, $P_{Ci}$, and the level of heterogeneity (between-trial variance), $\tau^2$. The results from the survey of the Cochrane Heart Group meta-analyses and our subsequent considerations informed the values and distributions of each of the above simulation components.

***Distribution of trial sizes***

Figure S13 and the 'trial sample size' columns in Table S1 in the supporting information provide an overview of the distribution of trial sample sizes among the surveyed Cochrane Heart Group mortality meta-analyses. After post hoc inspection of Figure S6 we calculated the proportion of trial sizes between 20, 200, 500, 1000, 2000, 5000, 10,000, and 50,000 participants. The estimated proportions are shown in table S2 in the supporting information. For our simulations, we decided only to simulate trials smaller than 5000 patients. We did this because intervention effects from large trials are assumed to come close to the overall average effect, whereas larger variations between intervention effects are typically observed across smaller trials.[25,26] Thus, the random-effects distributional assumptions, such as large between-trial variance, that allow for

very large trials to yield vastly different intervention effect estimates do not seem to be representative of meta-analyses in general.

Because larger trials are still rare in many medical areas, we decided to add a distribution of trial sizes to our simulations that more adequately reflected the spectrum and distribution of trial sample sizes that have been reported in the literature and seemed consistent with our own experience with meta-analysis in systematic reviews.[27,28] We decided on a second trial sample size distribution under which there is an 80% chance that a trial will have a sample size between 20 and 200 patients and a 20% chance that a trial will have a sample size between 200 and 500 patients. We have referred to this as a 'common' distribution of trial sample sizes in our paper.

***Distribution of control group risks***

In the survey of the Cochrane Heart Group meta-analyses, the median control group risks were generally lower than 10%. Based on table S1 in the supporting information, we assumed two distributions for the 'true' trial control group risk: 'low' and 'moderately low'. For low risks we assumed that the underlying trial control group risk would follow a uniform distribution between 1% and 5% (average of 3%). For moderately low risks, we assumed the underlying trial control group risk would follow a uniform distribution between 5% and 15% (average of 10%).

Subsequent considerations led us to additionally consider higher control group risks to increase generalizability of our results. In particular, we assumed that a 'moderate' control group risk could be represented by a uniform distribution between 15% and 40% (average of 27.5%), and

that a 'high' control group risk could be represented by a uniform distribution between 40% and 80% (average of 60%).

***Between-trial variance***

In the Cochrane Heart Group survey, the between-trial variance (heterogeneity) estimates spanned from 0.00 to 0.16, with the most common values being either truncated at 0.00 or in the interval of 0.03 to 0.07. The DerSimonian-Laird estimator is, however, known to underestimate the between-trial variance.[29,30] For this reason, we picked the three between-trial variance values 0.05, 0.15, and 0.25, which were moderately larger than the DerSimonian-Laird estimates observed in the Cochrane Heart Group meta-analysis sample. We believe that these values cover the spectrum in which most between-trial variance estimates fall, and thus, we did not add additional simulation values to those inferred from the survey.

***Underlying true intervention effect***

We considered two hypothetical situations: one where no underlying intervention effect exists (RRR = 0%), and one where a small but possibly unimportant intervention effect exists (RRR = 10%). Of note, the selected underlying true intervention effects were chosen to fit the objectives of this study, but not the results of the Cochrane Heart Group survey.

**The simulation setup**

First, we drew, with probabilities, the interval from which the trial sample size was to be sampled (Table S2 for the Cochrane Heart Group based trial sample sizes, or as given above for `common` trial sample sizes). We then drew the trial sample size, n, from a uniform distribution

on the interval that corresponded to the trial size category. The number of patients recruited to each intervention arm was set equal to n/2 (rounded up if n was an odd number).

We drew the trial specific control group risk, $P_{Ci}$, from a uniform distribution on one of the intervals given above (corresponding to the given scenario), and subsequently drew the number of observed events in the control group from a binomial distribution $e_{iC} \sim bin(n_i, P_{Ci})$.
We drew the underlying true trial intervention effects as log relative risks from a normal distribution, $ln(RR_i) \sim N(\mu, \tau^2)$, where $\mu$ is the natural logarithm of the underlying 'true' relative risk. Lastly, we drew the observed number of events in the intervention group from a binomial distribution $e_{iE} \sim bin(n_i, P_{Ei})$, where $P_{Ei} = P_{Ci} \cdot RR_i$. For each scenario, we simulated 20,000 meta-analysis data sets, each including 100 trials.

## Tables and Figures

| Scenario parameters | | | | Number of patients required for the probability of overestimation to drop below | | | Number of events required for the probability of overestimation to drop below | | |
|---|---|---|---|---|---|---|---|---|---|
| ***True effect*** | ***Overestimation*** | ***PC*** | $\tau^2$ | **10%** | **5%** | **1%** | **10%** | **5%** | **1%** |
| *RRR = 0%* | *RRR > 30%* | 1%-5% | 0.05 | 2000 | 3500 | 8000 | 100 | 150 | 300 |
| | | | 0.15 | 2500 | 4500 | 10500 | 100 | 150 | 350 |
| | | | 0.25 | 3000 | 5500 | 11500 | 150 | 200 | 350 |
| | | 5%-15% | 0.05 | 1000 | 1500 | 3500 | 100 | 150 | 350 |
| | | | 0.15 | 1500 | 2500 | 6500 | 150 | 250 | 600 |
| | | | 0.25 | 1500 | 3500 | 8000 | 200 | 350 | 750 |
| | *RRR > 20%* | 1%-5% | 0.05 | 5500 | 9000 | 19500 | 200 | 300 | 600 |
| | | | 0.15 | 6500 | 10500 | 21500 | 250 | 350 | 650 |
| | | | 0.25 | 6500 | 11500 | 23000 | 250 | 350 | 700 |
| | | 5%-15% | 0.05 | 2500 | 4000 | 9000 | 200 | 400 | 850 |
| | | | 0.15 | 3000 | 6500 | 13000 | 350 | 600 | 1250 |
| | | | 0.25 | 4500 | 8000 | 16500 | 450 | 750 | 1650 |
| *RRR = 10%* | *RRR > 30%* | 1%-5% | 0.05 | 4000 | 7000 | 14500 | 150 | 250 | 450 |
| | | | 0.15 | 5500 | 9000 | 18000 | 200 | 300 | 550 |
| | | | 0.25 | 5500 | 9000 | 18500 | 200 | 300 | 550 |
| | | 5%-15% | 0.05 | 2000 | 3000 | 7500 | 200 | 300 | 700 |
| | | | 0.15 | 2500 | 5500 | 11000 | 250 | 450 | 1000 |
| | | | 0.25 | 3500 | 7000 | 14000 | 350 | 600 | 1250 |
| | *RRR > 20%* | 1%-5% | 0.05 | 16500 | 26500 | >50000 | 500 | 800 | 1650 |
| | | | 0.15 | 15000 | 25000 | >50000 | 500 | 800 | 1500 |
| | | | 0.25 | 14500 | 24000 | >50000 | 450 | 750 | 1450 |
| | | 5%-15% | 0.05 | 7500 | 13500 | 26500 | 700 | 1250 | 2500 |
| | | | 0.15 | 10000 | 17000 | 37000 | 950 | 1600 | 3400 |
| | | | 0.25 | 12000 | 19500 | 40000 | 1150 | 1850 | 3750 |

**Table 1** Presents the required number of patients and events for the probability of overestimation to drop below 10%, 5% and 1%, in scenarios where the control group risk is 'low' or 'moderately low' and where the distribution of trial sample sizes is based on a survey of 23 Cochrane Heart Group meta-analyses on mortality.

| Scenario parameters | | | OIS (required number of patients) | | | OIS (required number of events) | | |
|---|---|---|---|---|---|---|---|---|
| *Assumed effect* | *PC* | $I^2$ | β=20% | β=10% | β=5% | β=20% | β=10% | β=5% |
| *RRR = 30%* | 3% | 0% | 9600 | 13000 | 16000 | 250 | 350 | 400 |
| | | 25% | 13000 | 17000 | 21500 | 350 | 450 | 550 |
| | | 50% | 19500 | 26000 | 32000 | 500 | 650 | 800 |
| *RRR = 20%* | | 0% | 23000 | 30500 | 38000 | 600 | 850 | 1000 |
| | | 25% | 30500 | 41000 | 51000 | 850 | 1100 | 1350 |
| | | 50% | 46000 | 61000 | 76000 | 1250 | 1650 | 2050 |
| *RRR = 30%* | 10% | 0% | 2700 | 3600 | 4500 | 250 | 300 | 400 |
| | | 25% | 3500 | 5000 | 6000 | 300 | 400 | 500 |
| | | 50% | 5500 | 7500 | 9000 | 450 | 600 | 750 |
| *RRR = 20%* | | 0% | 6500 | 8500 | 10500 | 600 | 800 | 1000 |
| | | 25% | 8500 | 11500 | 14000 | 750 | 1000 | 1300 |
| | | 50% | 13000 | 17000 | 21500 | 1150 | 1550 | 1900 |
| *RRR = 30%* | 27.5% | 0% | 900 | 1100 | 1400 | 200 | 300 | 350 |
| | | 25% | 1100 | 1500 | 1800 | 250 | 350 | 450 |
| | | 50% | 1700 | 2200 | 2700 | 400 | 550 | 650 |
| *RRR = 20%* | | 0% | 1900 | 2600 | 3200 | 500 | 650 | 800 |
| | | 25% | 2600 | 3500 | 4300 | 650 | 850 | 1050 |
| | | 50% | 3900 | 5200 | 6400 | 950 | 1300 | 1600 |
| *RRR = 30%* | 60% | 0% | 200 | 300 | 400 | 150 | 200 | 200 |
| | | 25% | 300 | 400 | 500 | 200 | 250 | 300 |
| | | 50% | 500 | 600 | 800 | 250 | 350 | 400 |
| *RRR = 20%* | | 0% | 500 | 700 | 900 | 300 | 400 | 500 |
| | | 25% | 700 | 1000 | 1200 | 400 | 550 | 650 |
| | | 50% | 1100 | 1500 | 1800 | 600 | 800 | 950 |

**Table 2** Presents the calculated optimal information size (OIS) to detect RRR = 30% and RRR = 20% respectively depending on the underlying assumed control group risk (*PC*), a desired type I error of 5%, variations of the desired type II error ($\beta$ = 20%, 10%, or 5%) and the anticipated degree of heterogeneity. The re required number of events have been rounded up to the nearest number divisible by 50. The required number of patients have been rounded up to the nearest number divisible by 1000 when PC = 3% and PC = 10% and to the nearest number divisible by 100 when PC = 27.5% and PC = 60%.

| Simulation | | | | | Optimal Information Size (OIS) | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **PC** | **Overestimation** | **Acceptability** | **Patients** | **Events** | **PC** | **RRR** | **Power** | **Patients** | **Events** |
| 1%-5% | RRR > 30% | Good | 3500-5500 | 150-200 | 3% | 30% | 80% | 10000-20000 | 250-500 |
| | | Very Good | 7000-11500 | 250-350 | | | 90% | 13000-26000 | 350-650 |
| | | Excellent | 14500-18500 | 450-550 | | | 95% | 16000-32000 | 400-800 |
| | RRR > 20% | Good | 10000-15000 | 400-500 | | 20% | 80% | 23000-46000 | 600-1250 |
| | | Very Good | 20000-25000 | 600-800 | | | 90% | 30000-61000 | 850-1650 |
| | | Excellent | >50000 | 1400-1600 | | | 95% | 38000-76000 | 1000-2050 |
| 5%-15% | RRR > 30% | Good | 2000-4000 | 200-300 | 3% | 30% | 80% | 3000-5500 | 250-450 |
| | | Very Good | 3000-8000 | 300-700 | | | 90% | 3500-7500 | 300-600 |
| | | Excellent | 7000-14000 | 700-1200 | | | 95% | 4500-9000 | 400-750 |
| | RRR > 20% | Good | 7000-12000 | 600-1200 | | 20% | 80% | 6500-13000 | 600-1150 |
| | | Very Good | 9000-19000 | 1250-1850 | | | 90% | 8500-17000 | 800-1600 |
| | | Excellent | 26000-40000 | 2500-2800 | | | 95% | 10500-21000 | 1000-1900 |

**Table 3** Presents the comparison of the optimal information size to demonstrate a relevant intervention effect with the required number of patients and events to limit the risk of overestimation in simulation scenarios where the distribution of trial sample sizes was based on survey of Cochrane Heart Group meta-analyses.

| Simulation | | | | | Optimal information size | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **PC** | **Overestimation** | **Acceptability** | **Patients** | **Events** | **PC** | **RRR** | **Power** | **Patients** | **Events** |
| 1%-5% | RRR > 30% | Good | 2500 | 100 | 3% | 30% | 80% | 10-20000 | 250-500 |
| | | Very Good | 3500-4500 | 150-200 | | | 90% | 13-26000 | 350-650 |
| | | Excellent | 6000-7500 | 200-250 | | | 95% | 16-32000 | 400-800 |
| | RRR > 20% | Good | 4000-7500 | 150-250 | | 20% | 80% | 23-46000 | 600-1250 |
| | | Very Good | 7000-11000 | 250-400 | | | 90% | 30-61000 | 850-1650 |
| | | Excellent | 14000-19000 | 350-600 | | | 95% | 38-76000 | 1000-2050 |
| 5%-15% | RRR > 30% | Good | 1500 | 100-150 | 10% | 30% | 80% | 3000-5500 | 250-450 |
| | | Very Good | 2000-3000 | 200-250 | | | 90% | 3500-7500 | 300-600 |
| | | Excellent | 3500-4500 | 350-450 | | | 95% | 4500-9000 | 400-750 |
| | RRR > 20% | Good | 2500-3500 | 250-350 | | 20% | 80% | 6500-13000 | 600-1150 |
| | | Very Good | 4500-5500 | 450-600 | | | 90% | 8500-17000 | 800-1600 |
| | | Excellent | 11000-12000 | 900-1150 | | | 95% | 10500-21000 | 1000-1900 |
| 15%-40% | RRR > 30% | Good | 500-2500 | 150-700 | 27.5% | 30% | 80% | 800-1700 | 200-400 |
| | | Very Good | 1400-6200 | 400-1700 | | | 90% | 1100-2200 | 300-550 |
| | | Excellent | 4000-12000 | 1000-3000 | | | 95% | 1400-2700 | 350-650 |
| | RRR > 20% | Good | 1000-3000 | 300-850 | | 20% | 80% | 1900-3600 | 500-950 |
| | | Very Good | 2100-5400 | 550-1350 | | | 90% | 2600-5200 | 650-1300 |
| | | Excellent | 6200-11400 | 1500-2300 | | | 95% | 3200-6400 | 800-1600 |
| 40%-80% | RRR > 30% | Good | 200-1000 | 150-500 | 60% | 30% | 80% | 200-500 | 150-250 |
| | | Very Good | 600-1800 | 300-1000 | | | 90% | 300-600 | 200-350 |
| | | Excellent | 1100-3200 | 600-1600 | | | 95% | 400-800 | 200-400 |
| | RRR > 20% | Good | 700-3400 | 350-1950 | | 20% | 80% | 500-1100 | 300-600 |
| | | Very Good | 1400-5800 | 750-3500 | | | 90% | 700-1500 | 400-800 |
| | | Excellent | 4000-11000 | 2000-5000 | | | 95% | 900-1800 | 500-950 |

**Table 4** Presents the comparison of the optimal information size (OIS) to demonstrate a relevant intervention effect with the required number of patients and events to limit the risk of overestimation in simulation scenarios where the distribution of trial sample sizes was based on survey of Cochrane Heart Group meta-analyses.

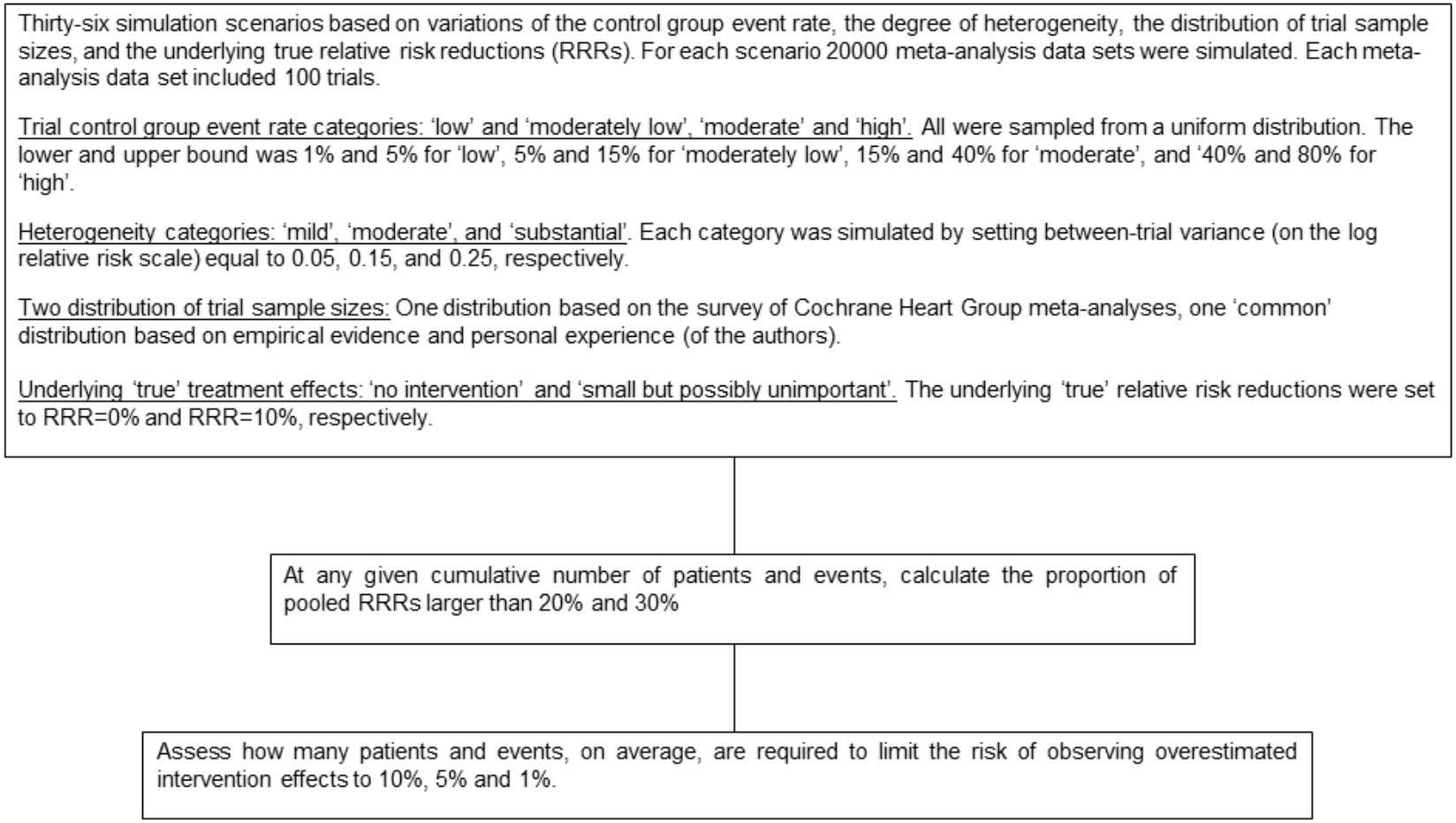



**Figure 1** Flowchart of simulations and analyses. Simulation scenarios that included combinations of Cochrane Heart Group survey based trial sample size distribution and either 'moderate' or 'high' control group risks were not performed.

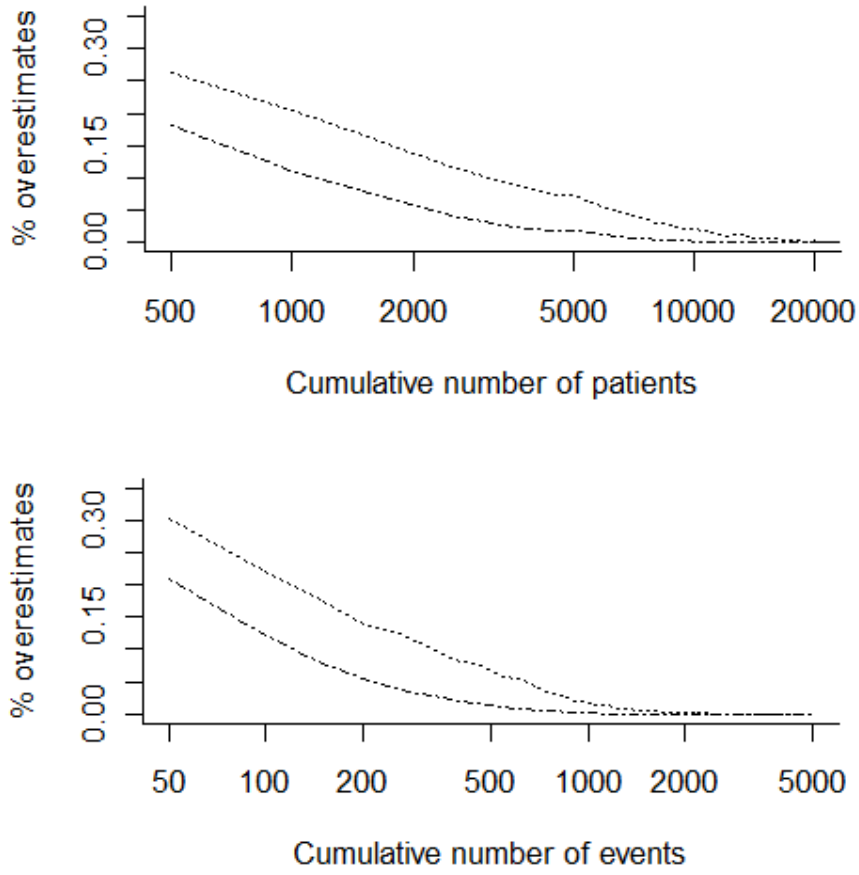



**Figure 2** Presents the proportions of pooled intervention effects overestimating the relative risk reduction with 30% (– · – ·) and 20% ( ····· ) in the scenario with no underlying intervention effect (i.e., RRR = 0%), the trial sample size distribution is based on the Cochrane Heart Group survey, the control group risk is moderate (i.e., drawn from a uniform distribution between 5% and 15%) and the heterogeneity is moderate (i.e., $\tau^2 = 0.15$). The proportion of pooled intervention effect estimates (the risk of overestimation) are plotted in relation to the cumulative number of patients (upper plot) and events (lower plot).

| | | MA sample size | | Trial Sample Size (#patients) | | | Trial control group event rate | | | Pooled RR | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **MA** | **Term** | #patients | #trials | Median | Quartile | Spectrum | Median | Quartiles | MinMax | **(95%CI)** | $\tau^2$ |
| 1 | Short | 2799 | 13 | 154 | 101-199 | 47-742 | 0.09 | 0.07-0.12 | 0.02-0.16 | 0.85(0.67-1.07) | 0.00 |
| 2 | Short | 7759 | 8 | 145 | 103-201 | 63-6800 | 0.05 | 0.04-0.06 | 0.01-0.35 | 0.99(0.93-1.06) | 0.00 |
| 3 | Short | 4832 | 15 | 238 | 163-337 | 85-1058 | 0.02 | 0.01-0.04 | 0.005-0.09 | 1.04(0.71-1.53) | 0.00 |
| 4 | Short | 3745 | 6 | 525 | 252-1022 | 62-1285 | 0.02 | 0.01-0.03 | 0.01-0.09 | 0.90(0.64-1.28) | 0.00 |
| 5 | Short | 2312 | 8 | 277 | 215-306 | 88-651 | 0.09 | 0.06-0.11 | 0.01-0.21 | 0.72(0.55-0.95) | 0.00 |
| 6 | Short | 2588 | 12 | 201 | 80-299 | 52-651 | 0.07 | 0.04-0.10 | 0.01-0.22 | 0.76(0.59-0.99) | 0.00 |
| 7 | Short | 2658 | 4 | 139 | 87-716 | 50-2330 | 0.40 | 0.35-0.43 | 0.17-0.48 | 0.84(0.64-1.09) | 0.03 |
| 9 | Short | 8412 | 21 | 164 | 102-311 | 26-3833 | 0.06 | 0.05-0.11 | 0.01-0.24 | 1.17(0.96-1.43) | 0.04 |
| 10 | Short | 6780 | 9 | 298 | 198-477 | 102-3833 | 0.09 | 0.06-0.18 | 0.01-0.22 | 1.18(0.90-1.55) | 0.07 |
| 11 | Short | 7473 | 14 | 187 | 150-329 | 26-3833 | 0.03 | 0.01-0.07 | 0.01-0.11 | 1.30(0.95-1.79) | 0.07 |
| 12 | Long | 2428 | 11 | 165 | 135-275 | 78-503 | 0.09 | 0.06-0.13 | 0.02-0.18 | 0.60(0.40-0.91) | 0.16 |
| 13 | Long | 1799 | 10 | 185 | 115-200 | 34-358 | 0.20 | 0.13-0.28 | 0.06-0.31 | 0.90(0.74-1.10) | 0.00 |
| 14 | Long | 5183 | 4 | 1209 | 355-2150 | 198-2268 | 0.11 | 0.06-0.18 | 0.06-0.20 | 0.98(0.85-1.13) | 0.00 |
| 15 | Long | 18679 | 6 | 3420 | 1217-5018 | 283-5522 | 0.07 | 0.05-0.09 | 0.04-0.17 | 1.00(0.92-1.09) | 0.00 |
| 17 | Long | 33201 | 15 | 360 | 211-1322 | 82-13406 | 0.04 | 0.02-0.11 | 0.01-0.22 | 0.87(0.73-1.03) | 0.03 |
| 18 | Long | 3604 | 6 | 219 | 156-728 | 88-2082 | 0.04 | 0.03-0.08 | 0.02-0.18 | 1.04(0.7-1.52) | 0.04 |
| 19 | Long | 10379 | 22 | 167 | 100-426 | 32-2481 | 0.07 | 0.04-0.14 | 0.01-0.34 | 0.77(0.61-0.99) | 0.12 |
| 20 | Long | 7546 | 11 | 441 | 209-816 | 98-2481 | 0.09 | 0.05-0.10 | 0.01-0.13 | 0.88(0.74-1-04) | 0.00 |
| 21 | Long | 12603 | 20 | 342 | 154-663 | 77-4165 | 0.34 | 0.22-0-45 | 0.08-0.82 | 0.64(0.58-0.71) | 0.02 |
| 22 | Long | 969 | 6 | 185 | 120-191 | 32-275 | 0.09 | 0.05-0.11 | 0.03-0.22 | 1.13(0.70-1.84) | 0.12 |

**Table S1** Presents the recorded meta-analysis and trial characteristics from the survey of Cochrane Heart Group mortality meta-analyses. The column labeled 'Quartile' contains the 25$^{th}$ to 75$^{th}$ percentile interval. The columns labeled 'Spectrum' contains the minimum and maximum value observed. The last column contains the DerSimonian-Laird estimate of the between-trial variance (on the log relative risk scale).

| Trial sample size interval | Estimated proportion for all trials | Estimated proportions excluding trials larger than 5000 patients | Proportions used for simulations |
|---|---|---|---|
| 20 to 200 | 50.8% | 52.3% | 50.0% |
| 201 to 500 | 26.6% | 27.4% | 27.5% |
| 501 to 1000 | 8.2% | 8.4% | 10.0% |
| 1001 to 2000 | 5.1% | 5.1% | 7.5% |
| 2001 to 5000 | 6.6% | 6.8% | 5.0% |
| 5000 to 10000 | 1.6% | - | - |
| 10000 to 50000 | 1.1% | - | - |

**Table S2** Estimated proportions of trial sample sizes based on the survey of Cochrane Heart Group meta-analysis as well as proportions used in our simulations.

| Scenario parameters | | | | Number of patients required for the probability of overestimation to drop below | | | Number of events required for the probability of overestimation to drop below | | |
|---|---|---|---|---|---|---|---|---|---|
| ***True effect*** | ***Overestimation*** | ***PC*** | $\tau^2$ | **10%** | **5%** | **1%** | **10%** | **5%** | **1%** |
| *RRR=0%* | *RRR>30%* | 15%-40% | 0.05 | 300 | 500 | 1000 | 100 | 150 | 250 |
| | | | 0.15 | 400 | 800 | 1500 | 150 | 250 | 400 |
| | | | 0.25 | 600 | 1000 | 2100 | 200 | 300 | 550 |
| | | 40%-80% | 0.05 | <100 | 200 | 600 | 100 | 150 | 350 |
| | | | 0.15 | 100 | 600 | 1200 | 200 | 350 | 650 |
| | | | 0.25 | 500 | 900 | 1800 | 300 | 450 | 1000 |
| | *RRR>20%* | 15%-40% | 0.05 | 700 | 1000 | 2100 | 200 | 300 | 550 |
| | | | 0.15 | 1000 | 1600 | 3300 | 300 | 450 | 900 |
| | | | 0.25 | 1300 | 2100 | 4200 | 400 | 600 | 1100 |
| | | 40%-80% | 0.05 | 400 | 700 | 1400 | 200 | 350 | 750 |
| | | | 0.15 | 800 | 1300 | 2600 | 450 | 700 | 1450 |
| | | | 0.25 | 1200 | 2100 | 4300 | 700 | 1100 | 2400 |
| *RRR=10%* | *RRR>30%* | 15%-40% | 0.05 | 500 | 900 | 1700 | 150 | 250 | 450 |
| | | | 0.15 | 800 | 1400 | 2600 | 250 | 350 | 650 |
| | | | 0.25 | 1100 | 1800 | 3400 | 300 | 450 | 900 |
| | | 40%-80% | 0.05 | 300 | 500 | 1100 | 200 | 300 | 600 |
| | | | 0.15 | 700 | 1100 | 2200 | 350 | 600 | 1200 |
| | | | 0.25 | 1000 | 1500 | 3200 | 500 | 800 | 1600 |
| | *RRR>20%* | 15%-40% | 0.05 | 1900 | 3100 | 6200 | 500 | 800 | 1500 |
| | | | 0.15 | 2500 | 4100 | 9800 | 700 | 1100 | 2150 |
| | | | 0.25 | 3500 | 5400 | 11500 | 850 | 1350 | 2300 |
| | | 40%-80% | 0.05 | 1100 | 1900 | 3900 | 650 | 1050 | 2000 |
| | | | 0.15 | 2300 | 4000 | 9700 | 1300 | 2200 | 4300 |
| | | | 0.25 | 3400 | 5800 | 11000 | 1950 | 3500 | >5000 |

**Table S3** Presents the required number of patients and events for the probability of overestimation to drop below 10%, 5% and 1%, in the simulation based on the sensitivity trial size distribution.

| Scenario parameters | | | | Number of patients required for the probability of overestimation to drop below | | | Number of events required for the probability of overestimation to drop below | | |
|---|---|---|---|---|---|---|---|---|---|
| ***True effect*** | ***Overestimation*** | ***PC*** | $\tau^2$ | **10%** | **5%** | **1%** | **10%** | **5%** | **1%** |
| *RRR=0%* | *RRR>30%* | 1%-5% | 0.05 | 1500 | 2500 | 4500 | 100 | 100 | 200 |
| | | | 0.15 | 1500 | 2500 | 4500 | 100 | 100 | 200 |
| | | | 0.25 | 1500 | 2500 | 4500 | 100 | 100 | 200 |
| | | 5%-15% | 0.05 | 1000 | 1500 | 2500 | 100 | 150 | 200 |
| | | | 0.15 | 1000 | 1500 | 2500 | 100 | 150 | 250 |
| | | | 0.25 | 1000 | 1500 | 3000 | 100 | 150 | 300 |
| | *RRR>20%* | 1%-5% | 0.05 | 3500 | 5000 | 10500 | 150 | 200 | 300 |
| | | | 0.15 | 3000 | 4500 | 9000 | 150 | 200 | 300 |
| | | | 0.25 | 2500 | 4000 | 7500 | 100 | 150 | 300 |
| | | 5%-15% | 0.05 | 1500 | 2500 | 4500 | 150 | 250 | 450 |
| | | | 0.15 | 2000 | 2500 | 5000 | 200 | 250 | 500 |
| | | | 0.25 | 2000 | 3000 | 5500 | 200 | 300 | 550 |
| *RRR=10%* | *RRR>30%* | 1%-5% | 0.05 | 2500 | 4000 | 7500 | 100 | 150 | 250 |
| | | | 0.15 | 2500 | 3500 | 6500 | 100 | 150 | 250 |
| | | | 0.25 | 2500 | 3500 | 6000 | 100 | 150 | 200 |
| | | 5%-15% | 0.05 | 1500 | 2000 | 3500 | 150 | 200 | 350 |
| | | | 0.15 | 1500 | 2500 | 4000 | 150 | 200 | 400 |
| | | | 0.25 | 1500 | 2500 | 4500 | 150 | 250 | 450 |
| | *RRR>20%* | 1%-5% | 0.05 | 7500 | 11500 | 19000 | 250 | 400 | 650 |
| | | | 0.15 | 5000 | 8500 | 17000 | 200 | 300 | 450 |
| | | | 0.25 | 4500 | 7000 | 14000 | 200 | 250 | 350 |
| | | 5%-15% | 0.05 | 3500 | 5500 | 11500 | 350 | 600 | 1150 |
| | | | 0.15 | 3500 | 5500 | 12000 | 350 | 550 | 900 |
| | | | 0.25 | 3500 | 5000 | 11000 | 350 | 550 | 900 |

**Table S4** Presents the required number of patients and events for the probability of overestimation to drop below 10%, 5% and 1%, in the simulation based on the sensitivity trial size distribution.

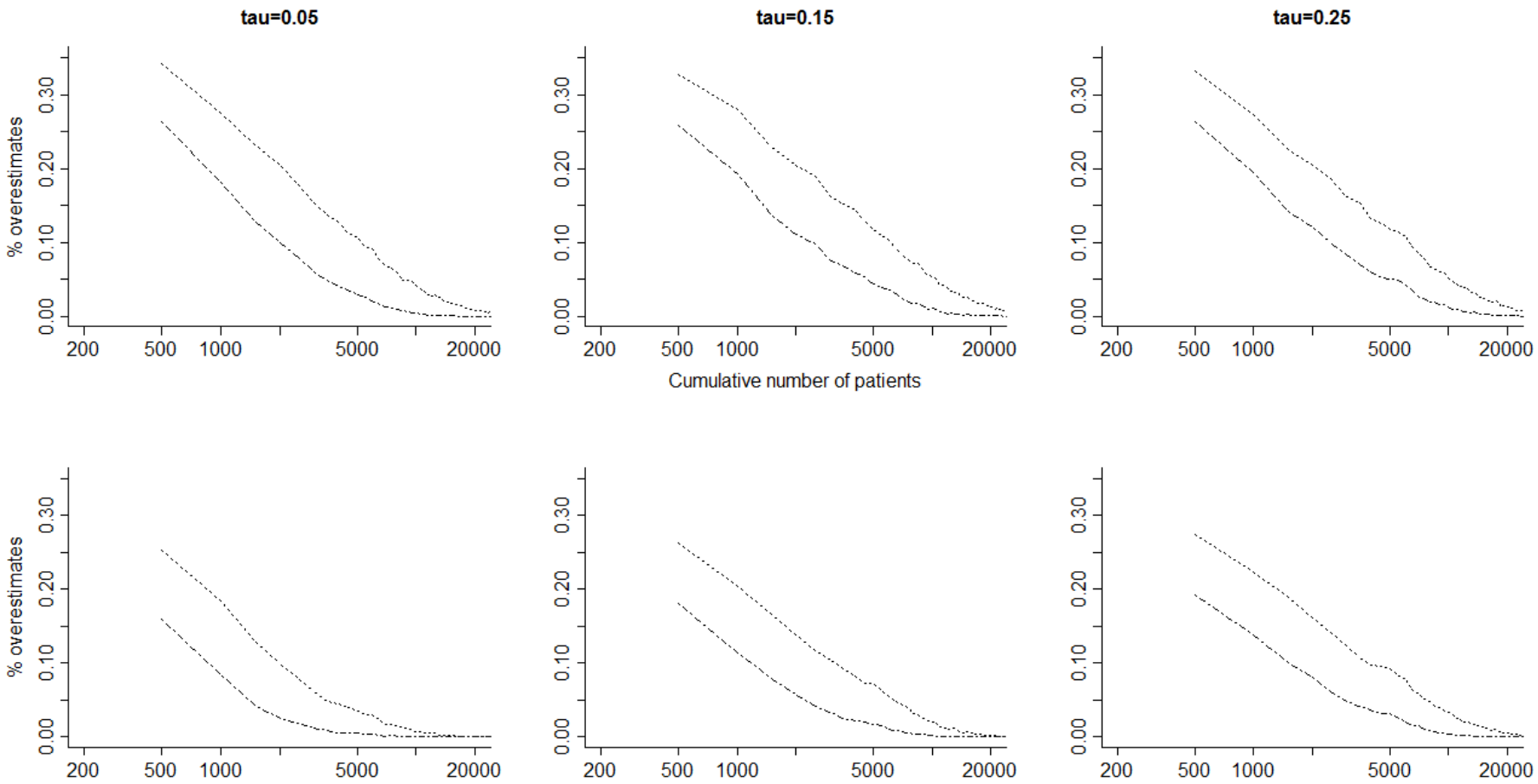



**Figure S1** Presents the proportions of pooled intervention effects exceeding a relative risk reduction of 30% (– · – ·) and 20% ( ····· ) when there is no underlying intervention effect (i.e., RRR = 0%), and where the distribution trial sample sizes are based on the survey of 23 Cochrane Heart Group meta-analyses. The proportions are plotted in relation to the cumulative number of patients. The upper three plots present the results from the simulated scenarios where the underlying 'true' trial control group risks are drawn from a uniform distribution between 1% and 5% ('low' risk), and the lower three plots present the results from the simulated scenarios where they are drawn from a uniform distribution between 5% and 15% ('moderately low' risk). The two left plots present results from scenarios with 'mild' heterogeneity ($\tau^2 = 0.05$), the middle two results from scenarios with moderate heterogeneity($\tau^2 = 0.15$), and the two right plots results from scenarios with substantial heterogeneity ($\tau^2 = 0.25$).

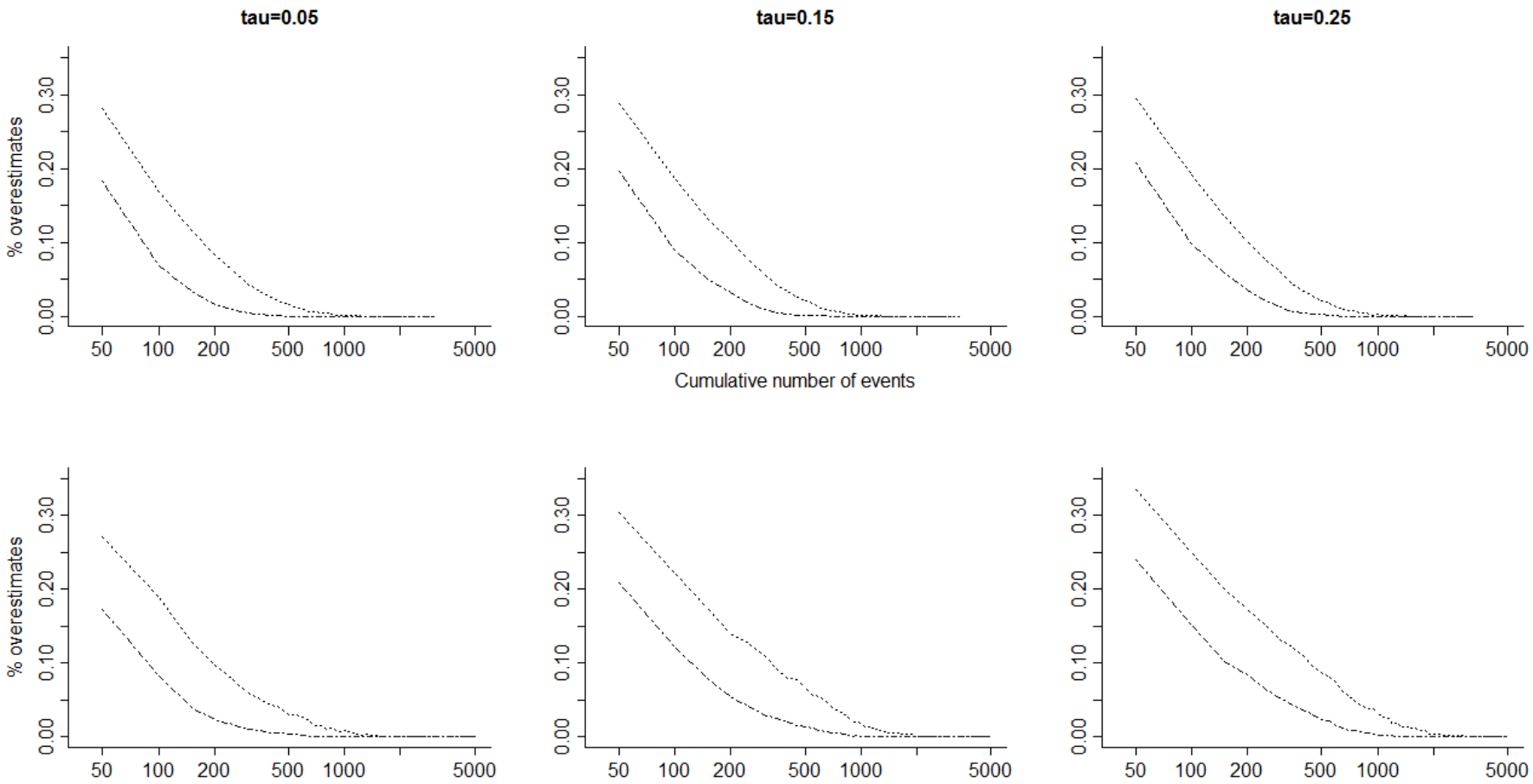



**Figure S2** Presents the proportions of pooled intervention effects exceeding a relative risk reduction of 30% (– · – ·) and 20% (·····) when there is no underlying intervention effect (i.e., RRR=0%), and where the distribution trial sample sizes are based on the survey of 23 Cochrane Heart Group meta-analyses. The proportions are plotted in relation to the cumulative number of events. The upper three plots present the results from the simulated scenarios where the underlying 'true' trial control group risks are drawn from a uniform distribution between 1% and 5% ('low' risk), and the lower three plots present the results from the simulated scenarios where they are drawn from a uniform distribution between 5% and 15% ('moderately low' risk). The two left plots present results from scenarios with 'mild' heterogeneity ($\tau^2 = 0.05$), the middle two results from scenarios with moderate heterogeneity($\tau^2 = 0.15$), and the two right plots results from scenarios with substantial heterogeneity ($\tau^2 = 0.25$).

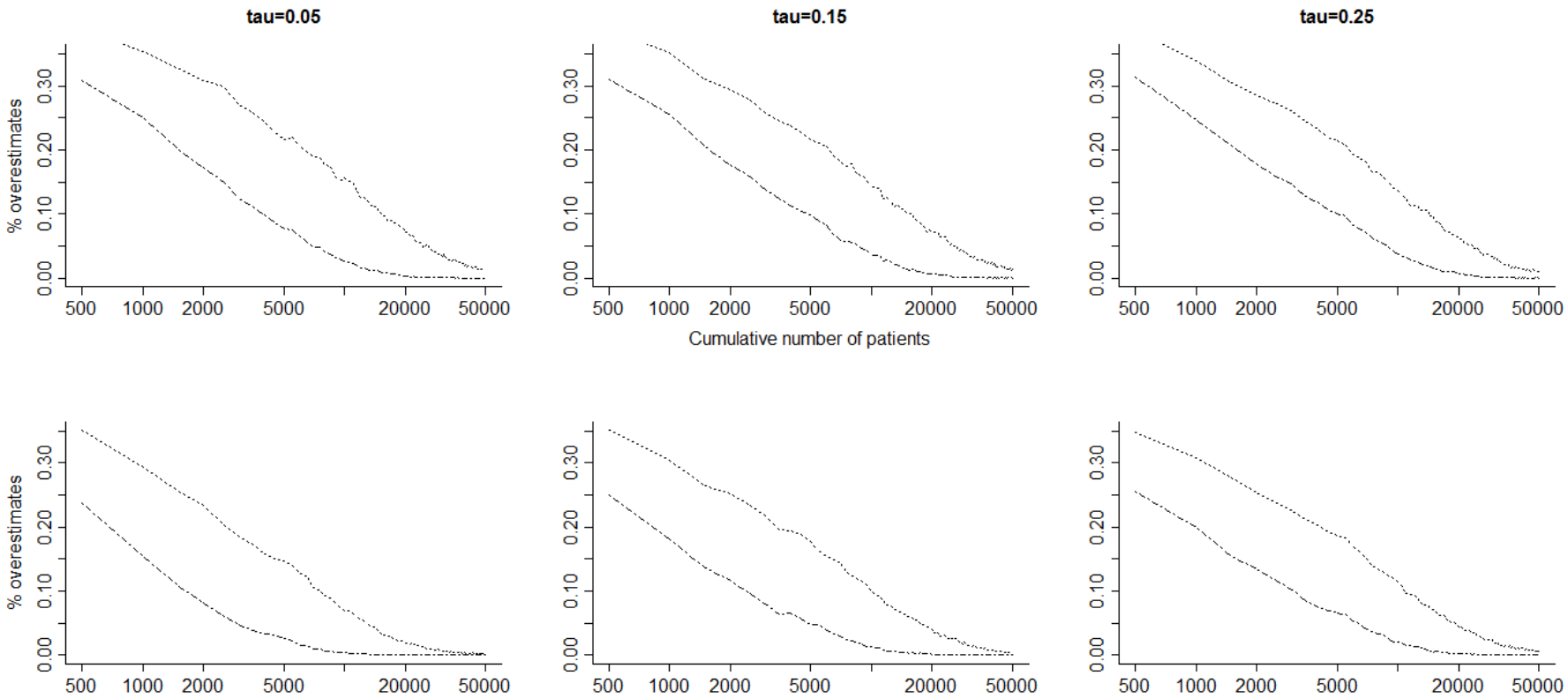



**Figure S3** Presents the proportions of pooled intervention effects exceeding a relative risk reduction of 30% (– · – ·) and 20% (····· ) when there is a small but potentially important intervention effect (i.e., RRR = 10%), and where the distribution trial sample sizes are based on the survey of 23 Cochrane Heart Group meta-analyses. The proportions are plotted in relation to the cumulative number of patients. The upper three plots present the results from the simulated scenarios where the underlying 'true' trial control group risks are drawn from a uniform distribution between 1% and 5% ('low' risk), and the lower three plots present the results from the simulated scenarios where they are drawn from a uniform distribution between 5% and 15% ('moderately low' risk). The two left plots present results from scenarios with 'mild' heterogeneity ($\tau^2 = 0.05$), the middle two results from scenarios with moderate heterogeneity ($\tau^2 = 0.15$), and the two right plots results from scenarios with substantial heterogeneity ($\tau^2 = 0.25$).

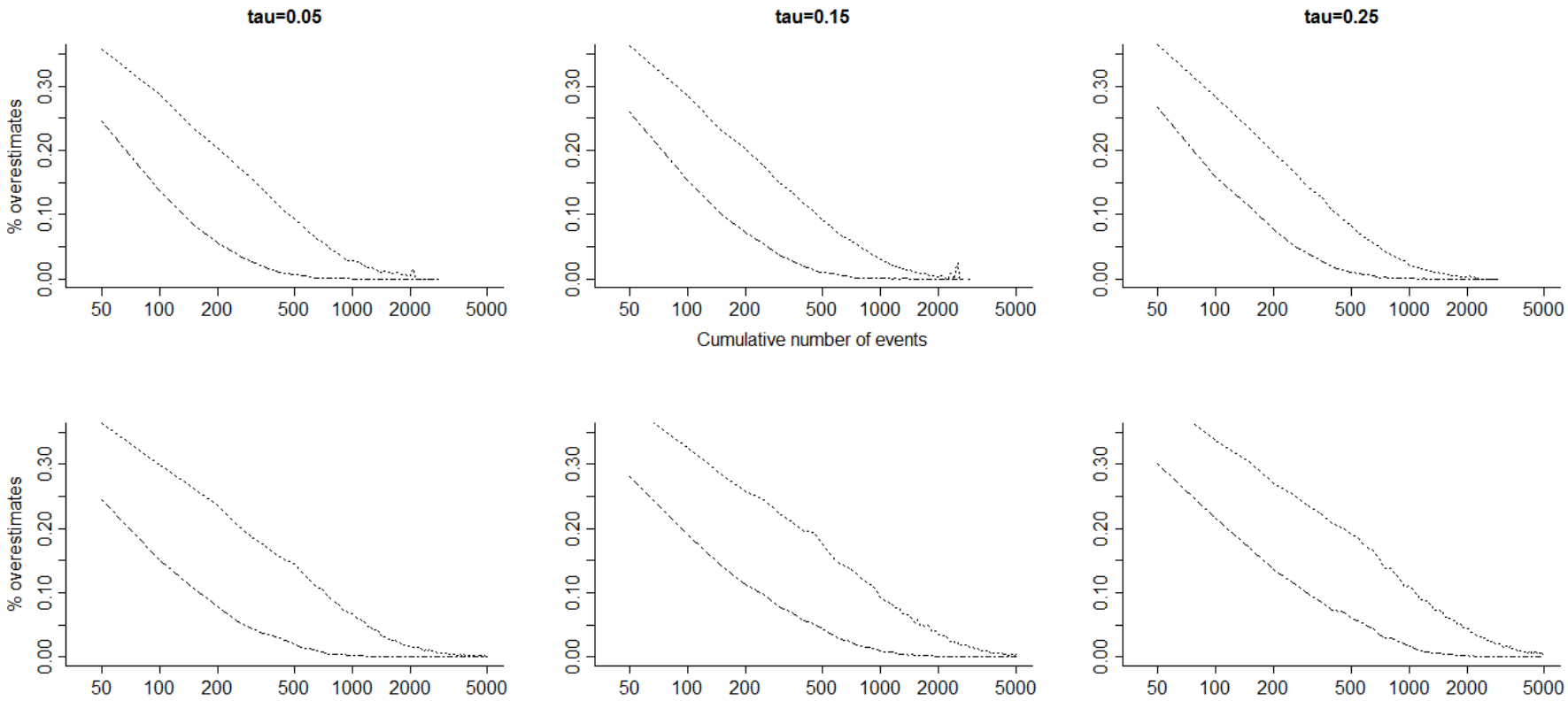



**Figure S4** Presents the proportions of pooled intervention effects exceeding a relative risk reduction of 30% (**– · – ·** ) and 20% (····· ) when there is small but potentially important intervention effect (i.e., RRR = 10%), and where the distribution trial sample sizes are based on the survey of 23 Cochrane Heart Group meta-analyses. The proportions are plotted in relation to the cumulative number of events. The upper three plots present the results from the simulated scenarios where the underlying 'true' trial control group risks are drawn from a uniform distribution between 1% and 5% ('low' risk), and the lower three plots present the results from the simulated scenarios where they are drawn from a uniform distribution between 5% and 15% ('moderately low' risk). The two left plots present results from scenarios with 'mild' heterogeneity ($\tau^2 = 0.05$), the middle two results from scenarios with moderate heterogeneity ($\tau^2 = 0.15$), and the two right plots results from scenarios with substantial heterogeneity ($\tau^2 = 0.25$).

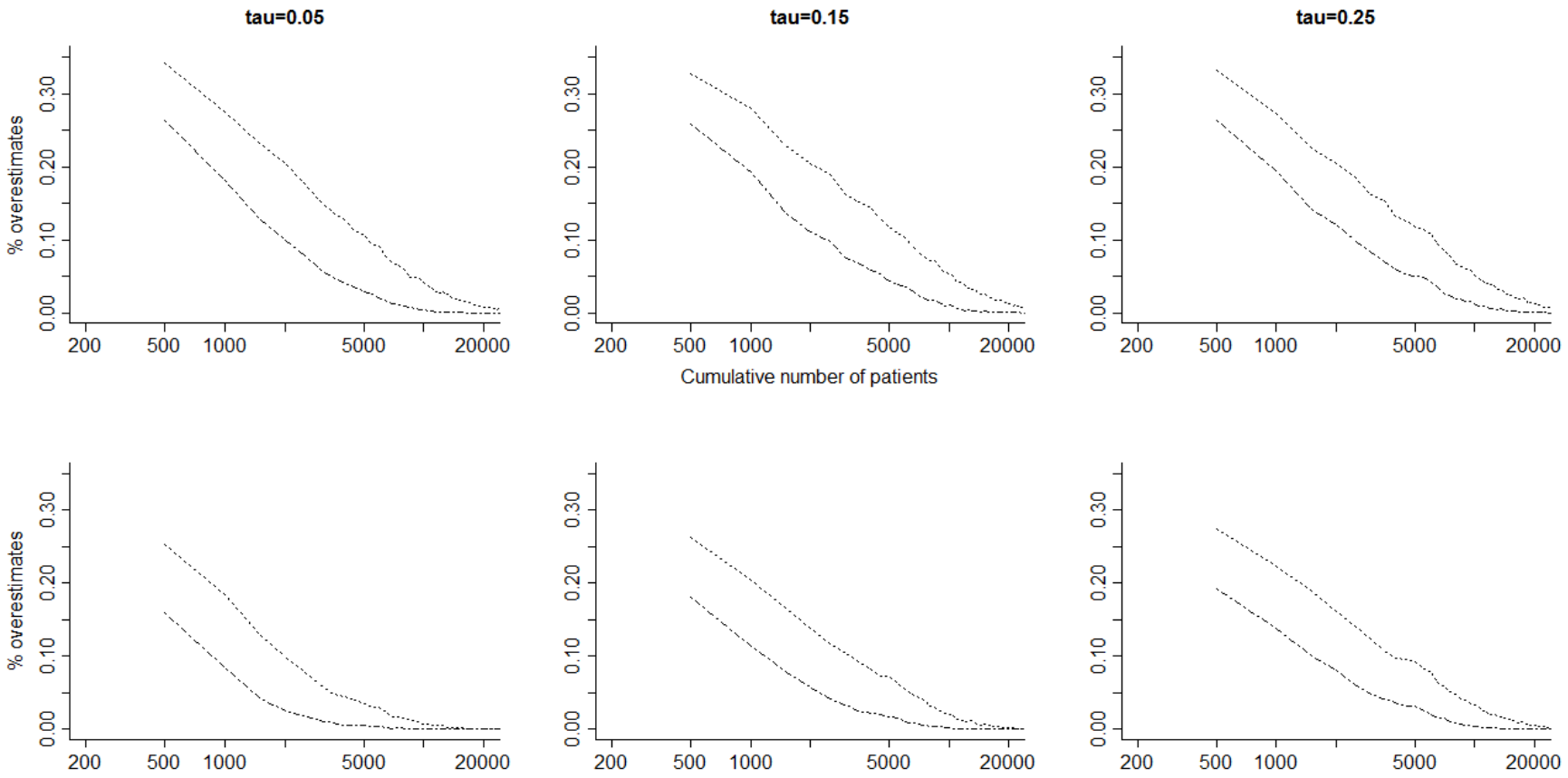



**Figure S5** Presents the proportions of pooled intervention effects exceeding a relative risk reduction of 30% (– · – ·) and 20% (····· ) when there is no underlying intervention effect (i.e., RRR = 0%), and where the distribution trial sample sizes are our assessment of what constitutes 'common' meta-analysis trial size distributions. The proportions are plotted in relation to the cumulative number of patients. The upper three plots present the results from the simulated scenarios where the underlying 'true' trial control group risks are drawn from a uniform distribution between 1% and 5% ('low' risk), and the lower three plots present the results from the simulated scenarios where they are drawn from a uniform distribution between 5% and 15% ('moderately low' risk). The two left plots present results from scenarios with 'mild' heterogeneity ($\tau^2 = 0.05$), the middle two results from scenarios with moderate heterogeneity ($\tau^2 = 0.15$), and the two right plots results from scenarios with substantial heterogeneity ($\tau^2 = 0.25$).

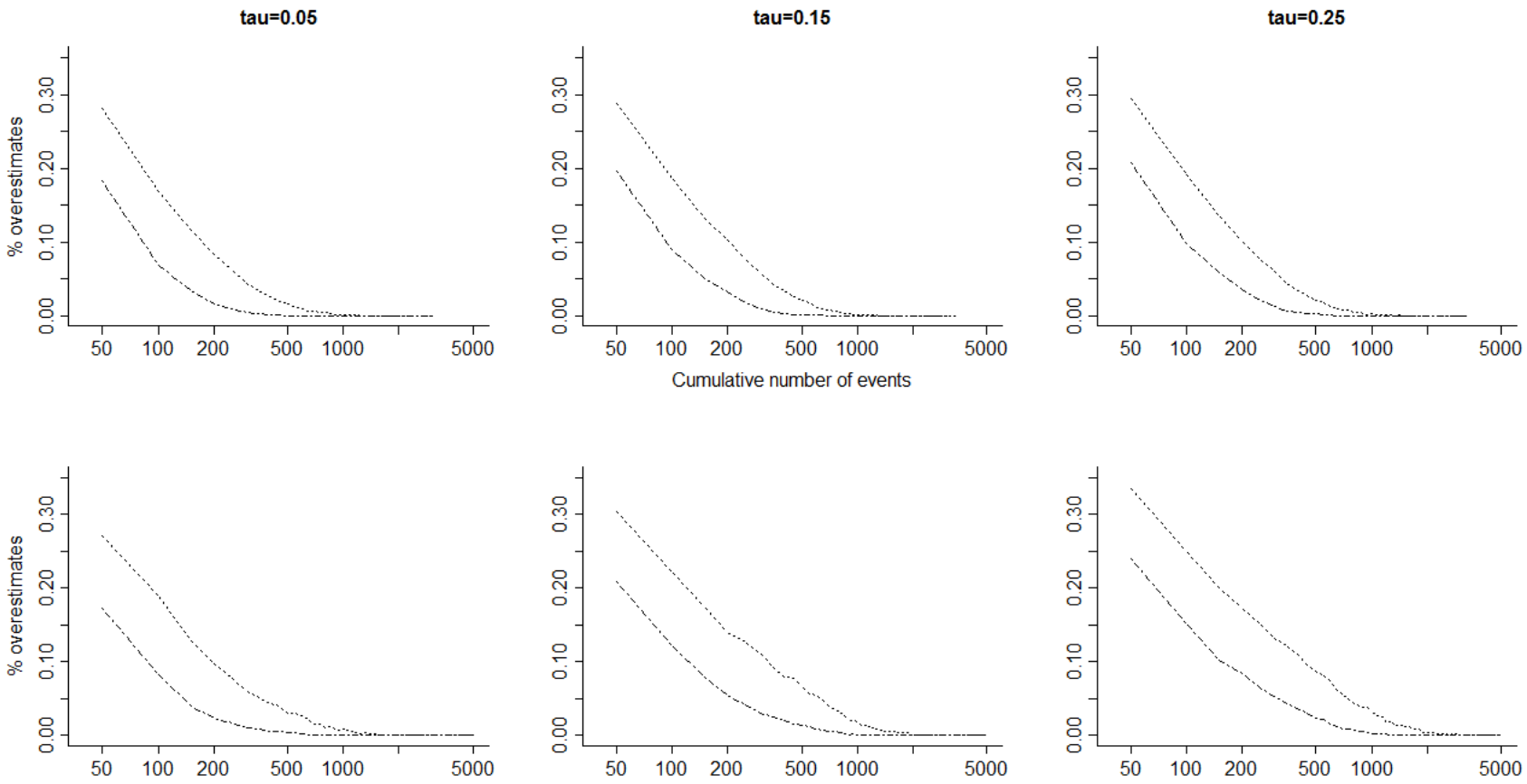



**Figure S6** Presents the proportions of pooled intervention effects exceeding a relative risk reduction of 30% (– · – · ) and 20% (····· ) when there is no underlying intervention effect (i.e., RRR = 0%), and where the distribution trial sample sizes are our assessment of what constitutes 'common' meta-analysis trial size distributions. The proportions are plotted in relation to the cumulative number of events. The upper three plots present the results from the simulated scenarios where the underlying 'true' trial control group risks are drawn from a uniform distribution between 1% and 5% ('low' risk), and the lower three plots present the results from the simulated scenarios where they are drawn from a uniform distribution between 5% and 15% ('moderately low' risk). The two left plots present results from scenarios with 'mild' heterogeneity ($\tau^2 = 0.05$), the middle two results from scenarios with moderate heterogeneity($\tau^2 = 0.15$), and the two right plots results from scenarios with substantial heterogeneity ($\tau^2 = 0.25$).

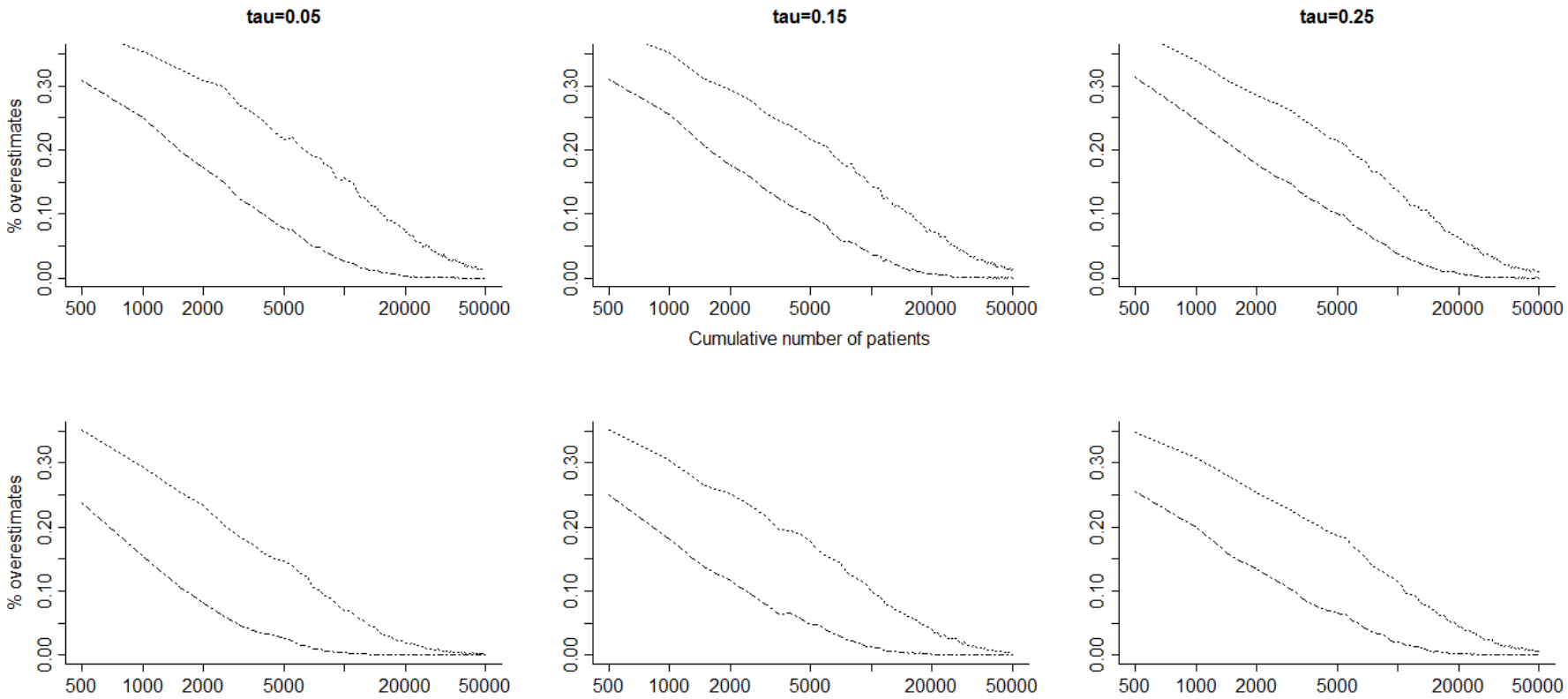



**Figure S7** Presents the proportions of pooled intervention effects exceeding a relative risk reduction of 30% (– · – ·) and 20% (·····) when there is small but potentially important intervention effect (i.e., RRR = 10%), and where the distribution trial sample sizes are our assessment of what constitutes 'common' meta-analysis trial size distributions. The proportions are plotted in relation to the cumulative number of patients. The upper three plots present the results from the simulated scenarios where the underlying 'true' trial control group risks are drawn from a uniform distribution between 1% and 5% ('low' risk), and the lower three plots present the results from the simulated scenarios where they are drawn from a uniform distribution between 5% and 15% ('moderately low' risk). The two left plots present results from scenarios with 'mild' heterogeneity ($\tau^2$ = 0.05), the middle two results from scenarios with moderate heterogeneity ($\tau^2$ = 0.15), and the two right plots results from scenarios with substantial heterogeneity ($\tau^2$ = 0.25).

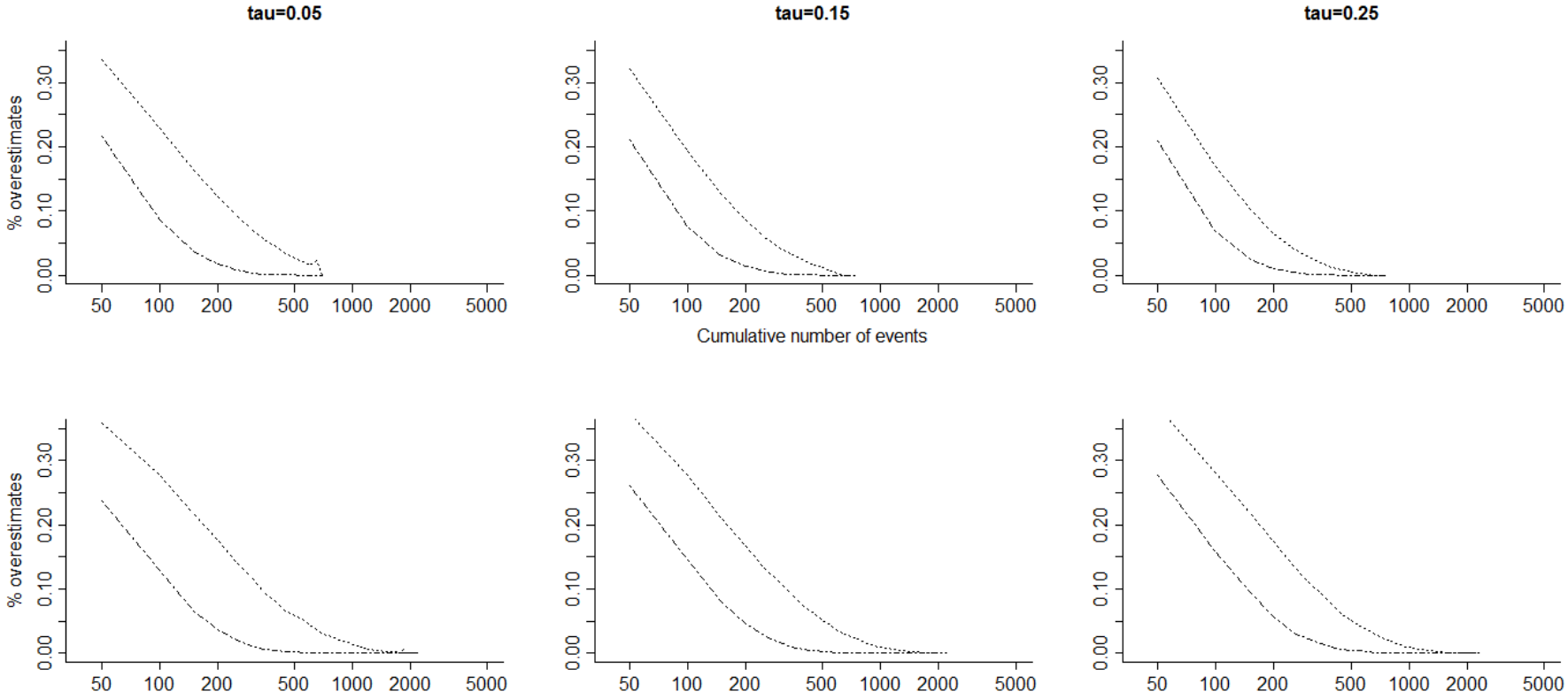



**Figure S8** Presents the proportions of pooled intervention effects exceeding a relative risk reduction of 30% (– · – ·) and 20% (·····) when there is small but potentially important intervention effect (i.e., RRR = 10%), and where the distribution trial sample sizes are our assessment of what constitutes 'common' meta-analysis trial size distributions. The proportions are plotted in relation to the cumulative number of events. The upper three plots present the results from the simulated scenarios where the underlying 'true' trial control group risks are drawn from a uniform distribution between 1% and 5% ('low' risk), and the lower three plots present the results from the simulated scenarios where they are drawn from a uniform distribution between 5% and 15% ('moderately low' risk). The two left plots present results from scenarios with 'mild' heterogeneity ($\tau^2$ = 0.05), the middle two results from scenarios with moderate heterogeneity ($\tau^2$ = 0.15), and the two right plots results from scenarios with substantial heterogeneity ($\tau^2$ = 0.25).

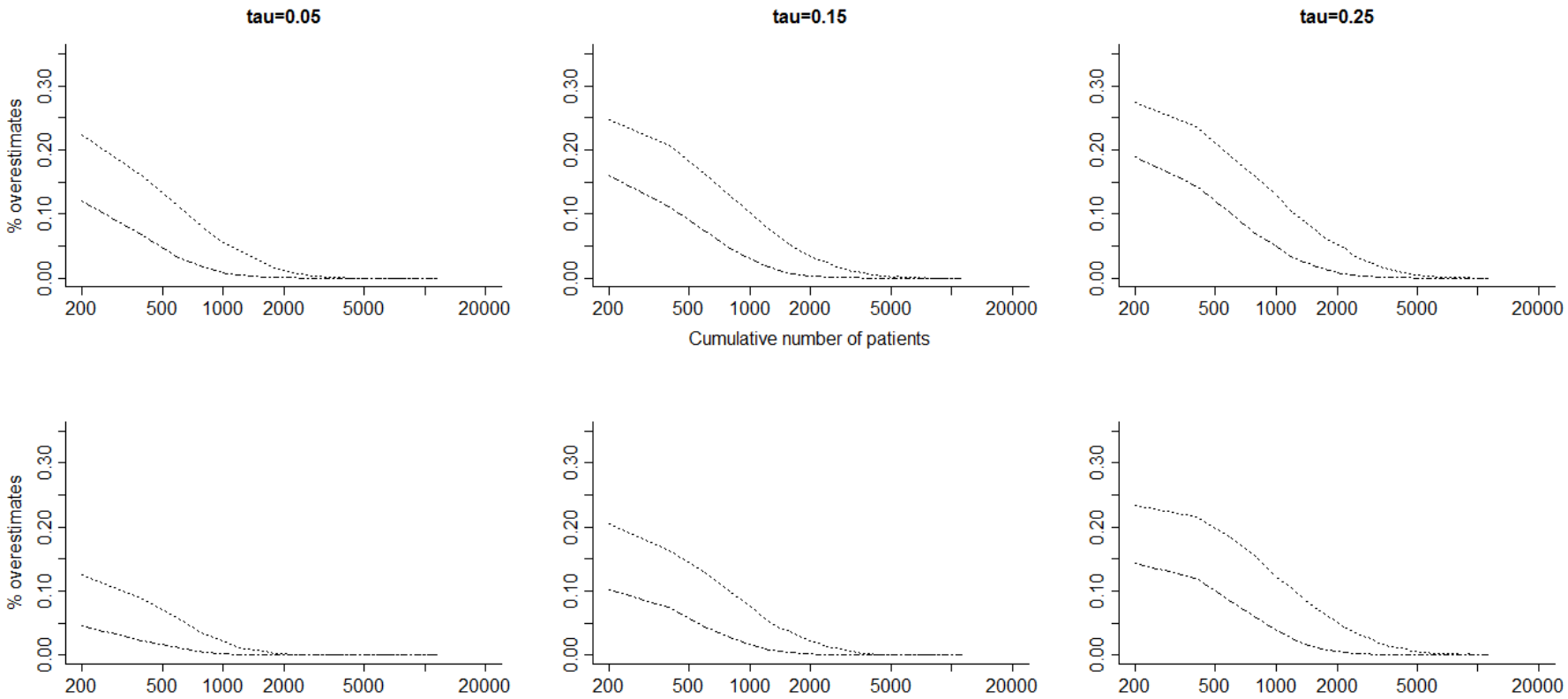



**Figure S9** Presents the proportions of pooled intervention effects exceeding a relative risk reduction of 30% (– · – ·) and 20% (·····) when there is no underlying intervention effect (i.e., RRR = 0%), and where the distribution trial sample sizes are our assessment of what constitutes 'common' meta-analysis trial size distributions. The proportions are plotted in relation to the cumulative number of patients. The upper three plots present the results from the simulated scenarios where the underlying 'true' trial control group risks are drawn from a uniform distribution between 15% and 40% ('moderate' risk), and the lower three plots present the results from the simulated scenarios where they are drawn from a uniform distribution between 40% and 80% ('high' risk). The two left plots present results from scenarios with 'mild' heterogeneity ($\tau^2 = 0.05$), the middle two results from scenarios with moderate heterogeneity ($\tau^2 = 0.15$), and the two right plots results from scenarios with substantial heterogeneity ($\tau^2 = 0.25$).

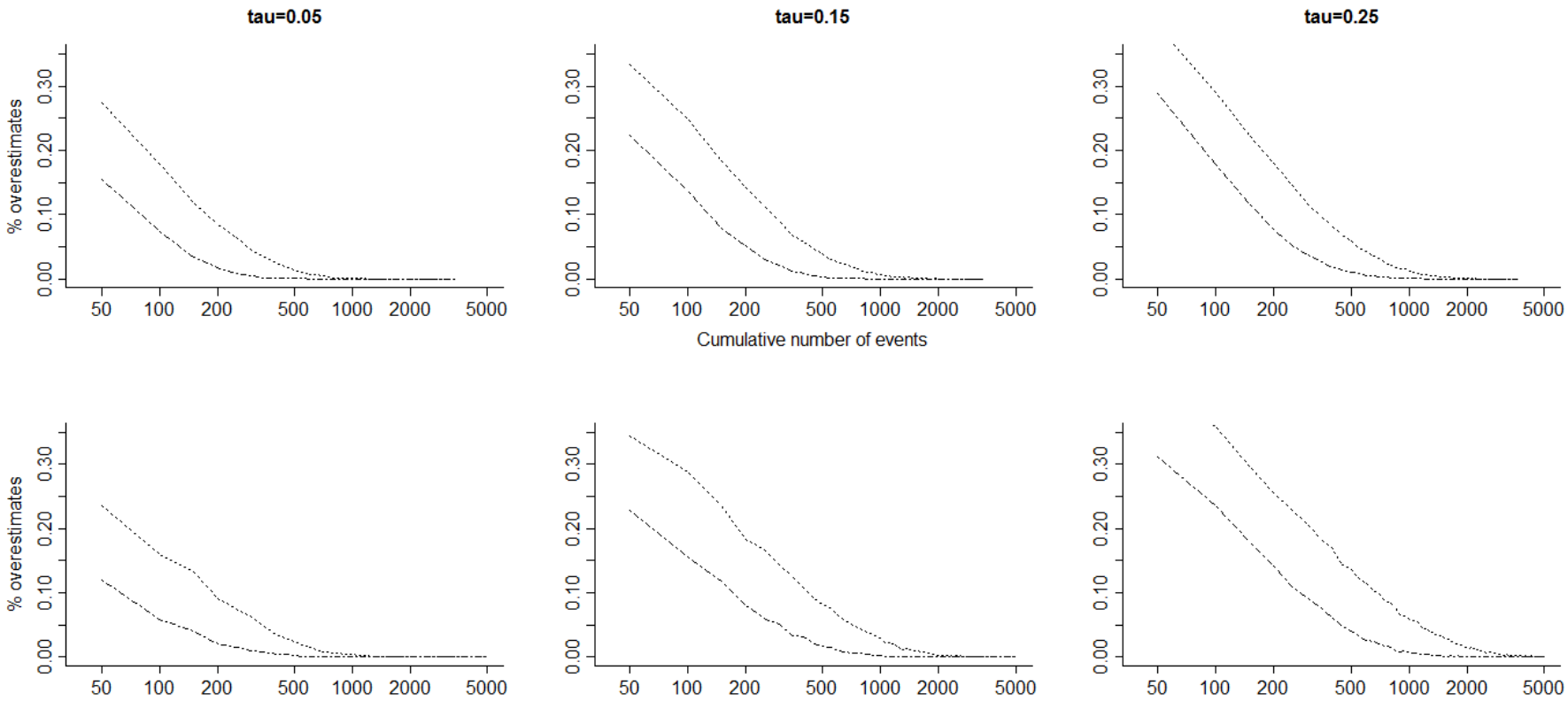



**Figure S10** Presents the proportions of pooled intervention effects exceeding a relative risk reduction of 30% (– · – · ) and 20% (····· ) when there is no underlying intervention effect (i.e., RRR = 0%), and where the distribution trial sample sizes are our assessment of what constitutes 'common' meta-analysis trial size distributions. The proportions are plotted in relation to the cumulative number of events. The upper three plots present the results from the simulated scenarios where the underlying 'true' trial control group risks are drawn from a uniform distribution between 15% and 40% ('moderate' risk), and the lower three plots present the results from the simulated scenarios where they are drawn from a uniform distribution between 40% and 80% ('high' risk). The two left plots present results from scenarios with 'mild' heterogeneity ($\tau^2 = 0.05$), the middle two results from scenarios with moderate heterogeneity ($\tau^2 = 0.15$), and the two right plots results from scenarios with substantial heterogeneity ($\tau^2 = 0.25$).

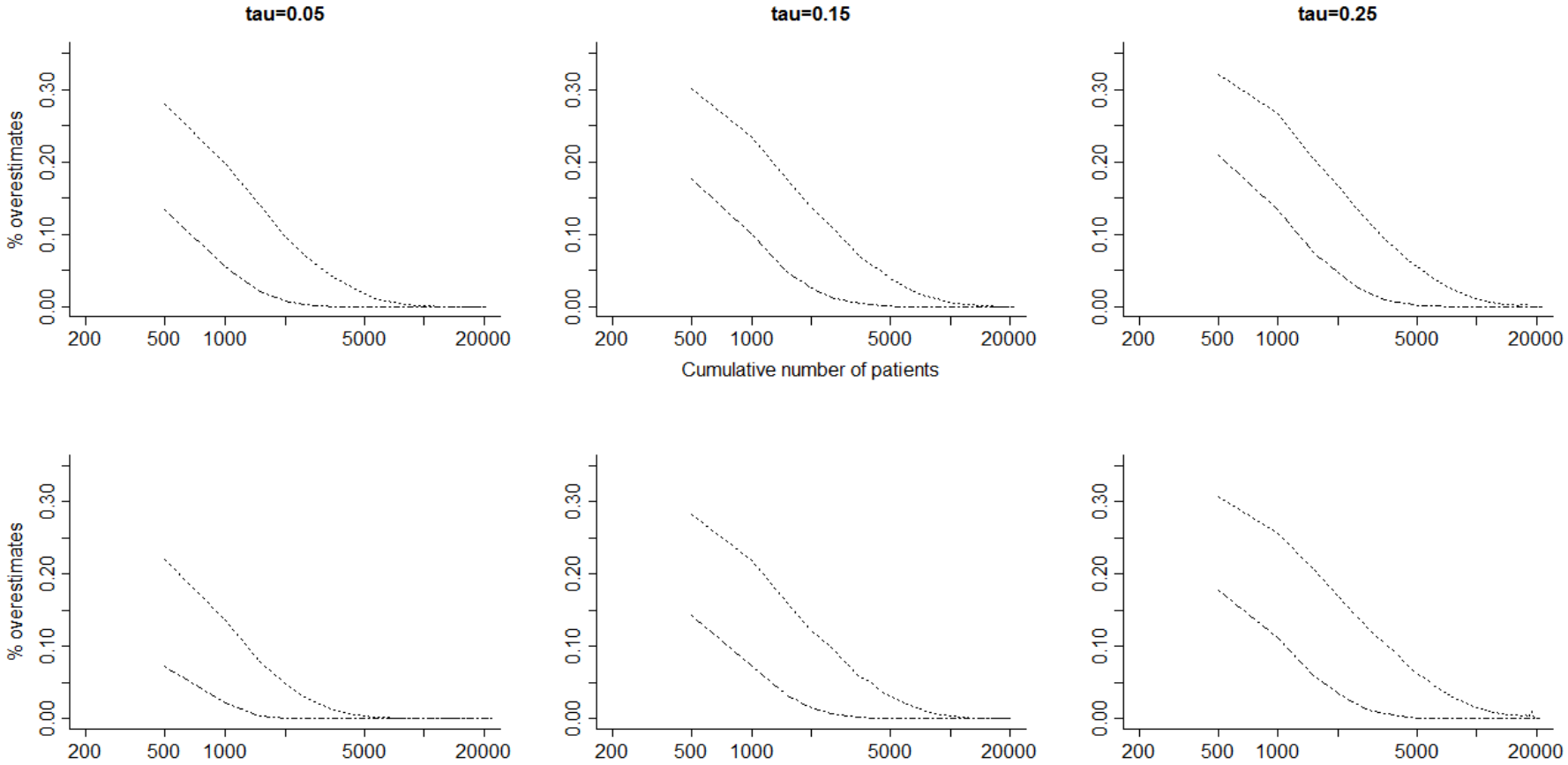



**Figure S11** Presents the proportions of pooled intervention effects exceeding a relative risk reduction of 30% (– · – · ) and 20% (····· ) when there is small but potentially important intervention effect (i.e., RRR = 10%), and where the distribution trial sample sizes are our assessment of what constitutes 'common' meta-analysis trial size distributions. The proportions are plotted in relation to the cumulative number of patients. The upper three plots present the results from the simulated scenarios where the underlying 'true' trial control group risks are drawn from a uniform distribution between 15% and 40% ('moderate' risk), and the lower three plots present the results from the simulated scenarios where they are drawn from a uniform distribution between 40% and 80% ('high' risk). The two left plots present results from scenarios with 'mild' heterogeneity

($\tau^2 = 0.05$), the middle two results from scenarios with moderate heterogeneity ($\tau^2 = 0.15$), and the two right plots results from scenarios with substantial heterogeneity ($\tau^2 = 0.25$).

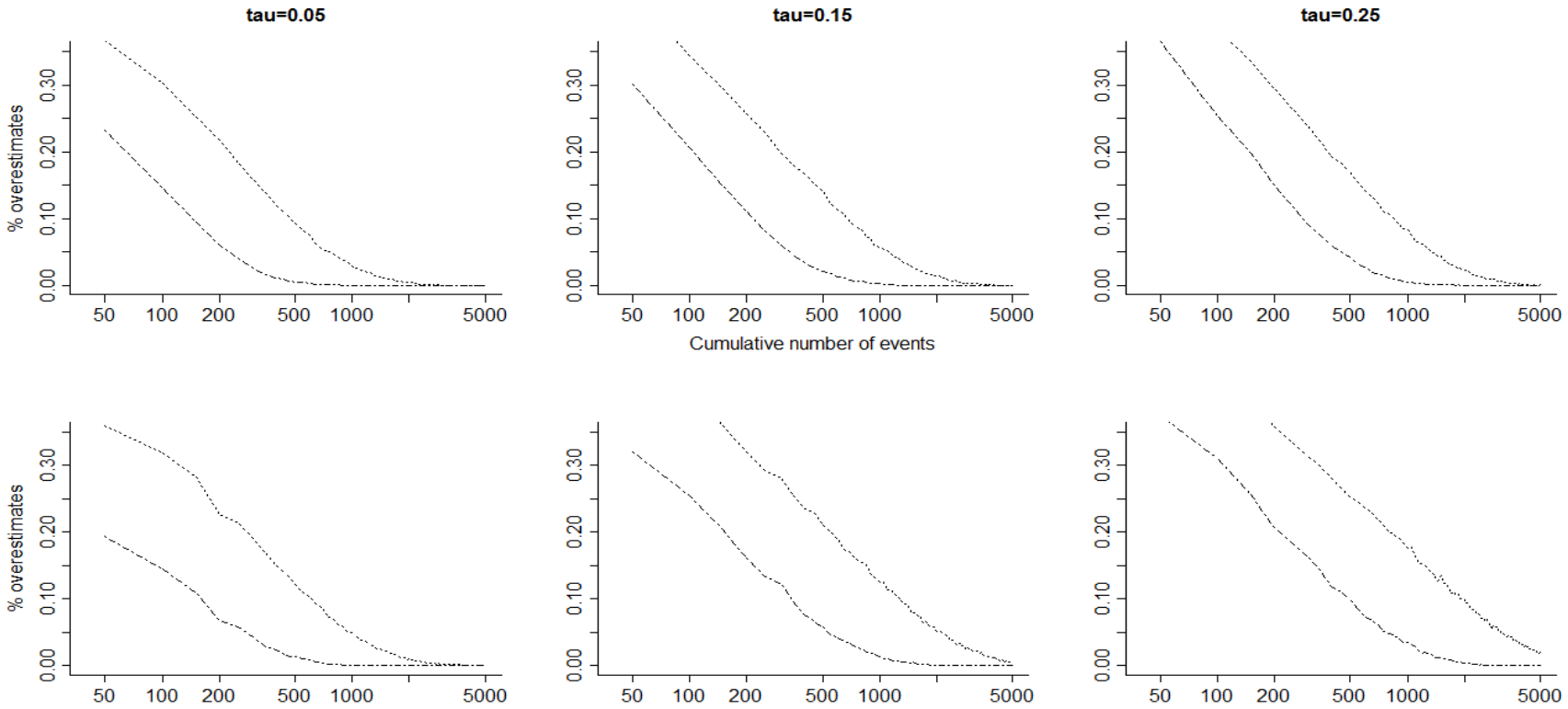



**Figure S12** Presents the proportions of pooled intervention effects exceeding a relative risk reduction of 30% (– · – ·) and 20% (·····) when there is small but potentially important intervention effect (i.e., RRR = 10%), and where the distribution trial sample sizes are our assessment of what constitutes 'common' meta-analysis trial size distributions. The proportions are plotted in relation to the cumulative number of events. The upper three plots present the results from the simulated scenarios where the underlying 'true' trial control group risks are drawn from a uniform distribution between 15% and 40% ('moderate' risk), and the lower three plots present the results from the simulated scenarios where they are drawn from a uniform distribution between 40% and 80% ('high' risk). The two left plots present results from scenarios with 'mild' heterogeneity ($\tau^2 = 0.05$), the middle two results from scenarios with moderate heterogeneity ($\tau^2 = 0.15$), and the two right plots results from scenarios with substantial heterogeneity ($\tau^2 = 0.25$).

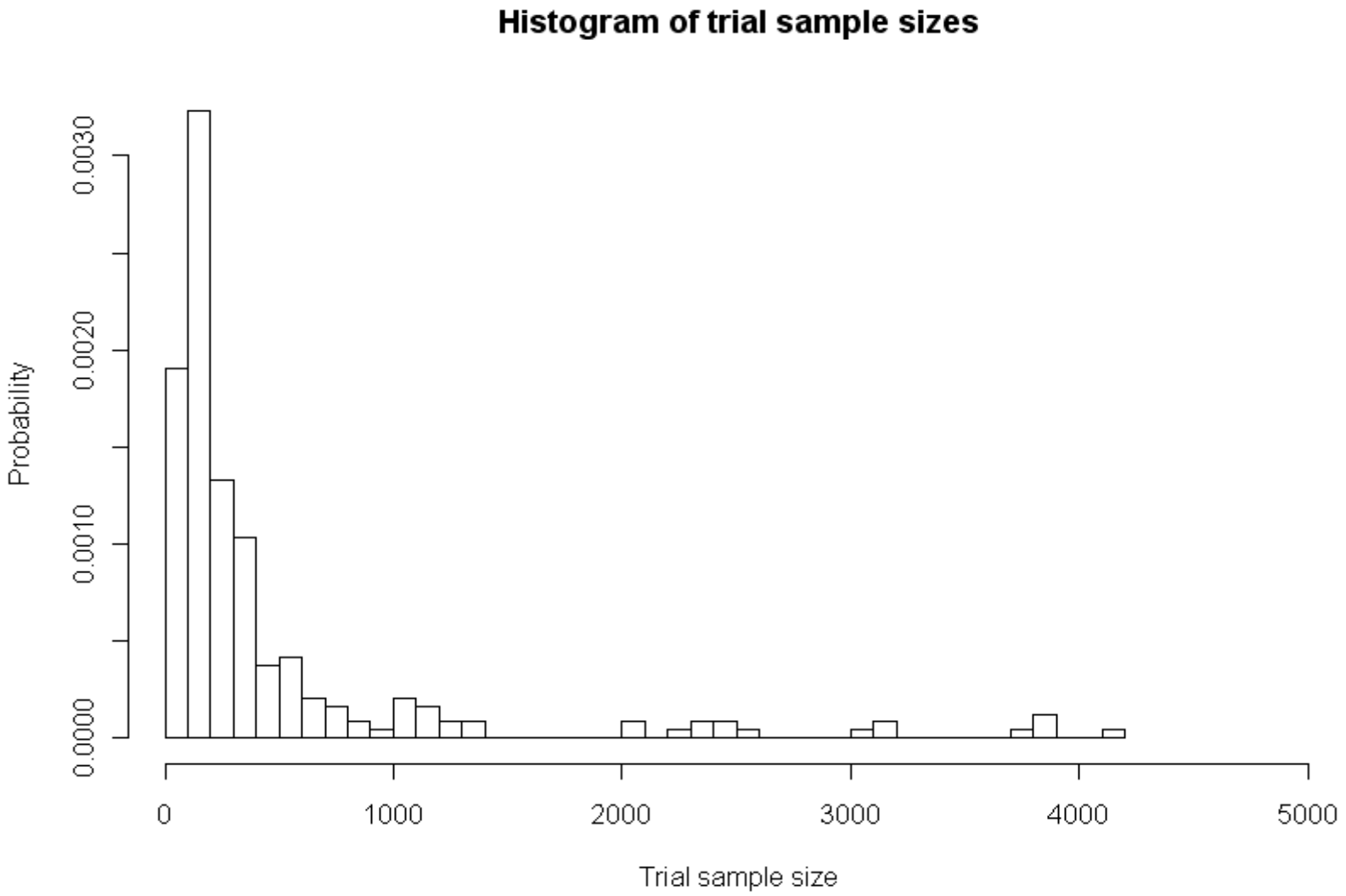



**Figure S13** Histogram of trial sample sizes in the surveyed Cochrane heart group meta-analyses

# Chapter 3: Evolution of heterogeneity ($I^2$) estimates and their 95% confidence intervals in large meta-analyses

Authors:

Kristian Thorlund

Georgina Imberger

Bradley C. Johnston

Michael Walsh

Tahany Awad

Lehana Thabane

Christian Gluud

PJ Devereaux

Jørn Wetterslev

Word count:

Abstract: 325

Manuscript: 3144

## Abstract

**Background:** Assessment of heterogeneity is essential in systematic reviews and meta-analyses of clinical trials. The most commonly used heterogeneity measure, $I^2$, provides an estimate of the proportion of variability in a meta-analysis that is explained by differences between the included trials rather than by sampling error. Recent studies have raised concerns about the reliability of $I^2$ estimates, due to their dependence on the precision of included trials and time-dependent biases. Authors have also advocated use of 95% confidence intervals (CIs) to express the uncertainty associated with $I^2$ estimates. However, no previous studies have explored how many trials and events are required to ensure stable and reliable $I^2$ estimates, or how 95% CIs perform as evidence accumulates.

**Methods and Findings:** To assess the stability and reliability of $I^2$ estimates and their 95% CIs, in relation to the cumulative number of trials and events in meta-analysis, we looked at 16 large Cochrane meta-analyses - each including a sufficient number of trials and events to reliably estimate $I^2$ - and monitored the $I^2$ estimates and their 95% CIs for each year of publication. In 10 of the 16 meta-analyses, the $I^2$ estimates fluctuated more than 40% over time. The median number of events and trials required before the cumulative $I^2$ estimates stayed within +/- 20% of the final $I^2$ estimate was 467 and 11. No major fluctuations were observed after 500 events and 15 trials. The 95% confidence intervals provided good coverage over time.

**Conclusion:** $I^2$ estimates need to be interpreted with caution when the meta-analysis only includes a limited number of events or trials. Confidence intervals for $I^2$ estimates provide good coverage as evidence accumulates, and are thus valuable for reflecting the uncertainty associated with estimating $I^2$.

## Introduction

Measures of heterogeneity are essential in systematic reviews and meta-analyses of clinical trials.[1-6] The most commonly used heterogeneity measure , $I^2$, provides an estimate of the proportion of variability in a meta-analysis that is explained by differences between the included trials rather than by sampling error.[2,3] Several studies have demonstrated important shortcomings of the $I^2$ measure.[7-12] $I^2$ estimates may be particularly unreliable in meta-analyses including a small number of trials (e.g., less than 10 trials) due to lack of power.[7,8] $I^2$ estimates may be underestimated as a result of time-lag bias.[9,10] Moreover, comparably higher or lower precision in the most recently added trials may inflate or deflate $I^2$ under different circumstances.[8,11]

Imprecise or biased estimates of heterogeneity can have serious consequences.[6,12] Underestimation of heterogeneity may inappropriately prevent exploration of the cause(s) of heterogeneity. Overestimation of heterogeneity may inappropriately prevent a meta-analysis actually being done. Overestimation may also trigger inappropriate exploration of the cause(s) of heterogeneity. For example, large $I^2$ estimates may prompt authors to exhaust all possibilities of subgroup analyses – a conduct notorious for its tendency to yield findings beyond replication.[13]

In response to the above identified shortcomings, it has been proposed that reported $I^2$ estimates should be accompanied by their associated 95% confidence interval (CI).[6,12] Confidence intervals may be a desirable addition to the single $I^2$ estimate; they give an appreciation of the spectrum of possible degrees of heterogeneity (e.g., mild to moderate), allowing for more appropriate interpretation of the overall intervention effect estimate. One concern, however, is the possibility that the $I^2$ estimate's dependence on power, trial weights, and time-lag bias may cause

fluctuations beyond the play of chance. With such fluctuations, the 95% CIs may not retain their desired coverage.

To explore these issues we selected a sample of 16 large Cochrane meta-analyses, each including a sufficient number of trials, patients and events to provide reliable treatment effect estimates and $I^2$ estimates. We retrospectively re-analysed the data for each meta-analysis, starting with the first chronological trial, and calculating a cumulative $I^2$ estimate and its associated 95% CI after each new trial was added to the meta-analysis. We then estimated the number of events and trials generally needed for $I^2$ estimates and 95% CIs to converge.

## Statistical framework and theoretical considerations

In this section we first outline the construct of the $I^2$ measure and its associated 95% CI. We secondly provide an overview of meta-analysis factors and properties of the $I^2$ measure that may inappropriately affect the magnitude of the $I^2$ estimate. Lastly, we provide the rationale for empirically studying $I^2$ estimate and their associated 95% CIs over time.

### Measuring heterogeneity between studies

Higgins et al. proposed the now widely popular measure of heterogeneity, $I^2$, as well as methods for calculating the associated 95% CIs.[2,3] $I^2$ expresses the proportion of variability in a meta-analysis which is explained by between-trial heterogeneity rather than by sampling error. Mathematically, $I^2$ is expressed as $I^2=\tau^2/(\sigma^2+\tau^2)$, where $\tau^2$ denotes the between-trial heterogeneity, $\sigma^2$ denotes some common sampling error across trials, and $\sigma^2+\tau^2$ is the total variation in the meta-analysis. $I^2$ is usually calculated as $(Q-df)/Qx100\%$, where $Q$ is the

Cochran's homogeneity test statistic and *df* is the degrees of freedom (the number of trials minus 1).[2,3,14] Higgins et al. explored a number of methods for obtaining 95% CIs of the $I^2$ estimate.[2] For this study, we will use the method referred to as the *test based* method in Higgins et al.[2] This method yields good coverage in most situations and is easy to calculate.[2] The required calculations for this method are outlined in the appendix.

**Factors affecting $I^2$ estimates**

$I^2$ estimates may be unreliable due to lack of power and precision[7,8,11] due to the presence of time-dependent biases,[9,10] or due to dependence on trial weights and precisions.

***Power and precision***

Since $I^2$ is a monotonically increasing function of Cochran's $Q$, large values for $Q$ result in large $I^2$ estimates and small values for $Q$ result in small $I^2$ estimates. The power of Cochran's $Q$ depends on the number of trials and the precision of the trials (i.e., the number of patients and events in the trials).[7,8,11] When the number of trials or their respective precision are small, Cochran's Q usually has inappropriately low power to detect heterogeneity, and therefore tends to yield conservative (low) test values.[7,8] To illustrate, the median number of trials is seven for Cochrane meta-analyses and 12 for meta-analyses published in paper journals.[15,16] The median sample size in randomized clinical trials is typically less than 100 in most medical specialties.[17,18] Thus, it is common for Cochran's Q to have low power. This lack of power is likely to cause underestimation of $I^2$, particularly if there are few events among the included trials.[7]

***Time-dependent bias***

Time-dependent bias (i.e., time-lag bias and publication bias) is known as a threat to the validity of the pooled estimate of effect in meta-analyses.[19-21] In addition, time-dependent bias may compromise the validity of heterogeneity estimates.[9,10] It is accepted that statistically significant trials with large intervention effect estimates usually get published the fastest.[21] If a meta-analysis is conducted at a time where all trials yield large promising treatment effects, the similarity across trials will result is a relatively small $I^2$ estimate. If the meta-analysis is updated some years later, this update is likely to include trials that found more moderate, neutral, or negative treatment effects. The inclusion of such trials will generate larger estimates of heterogeneity.

***Dependence on trial weights and precision***

From the mathematical expression $I^2=\tau^2/(\sigma^2+\tau^2)$, it is clear that relatively large sampling errors across trials will result in small $I^2$ estimates, and conversely, that relatively small sampling errors across trials will result in large $I^2$ estimates.[2,3,8,11] The "common" sampling error, $\sigma^2$, across trials may change considerably over time. For example, if early trials enroll a more homogeneous or heterogeneous set of patients than later trials, if they have shorter follow-up than later trials, or if changes are made to the definition of the outcome measure (e.g., the definition of myocardial infarction has changed considerably over the past decades); then the "common" sampling error may be considerably different at a later stage in a meta-analysis than it was in the early stage. Provided the between study variance, $\tau^2$, remains relatively stable over time, changes in the "common" sampling error may cause considerable changes in $I^2$ estimates over time. Further, if

the between-study variance incurs considerable changes over time, changes in the "common" sampling error may either inflate or deflate the representation of such changes through the $I^2$ estimate.

**The need to assess convergence *$I^2$ estimates and confidence intervals***

From the above discussion, it is evident that $I^2$ estimates may incur considerable fluctuations over time. Currently, no studies have explored the magnitude of this problem, and no recommendations exist as to how many events or trials are needed to achieve adequately stable $I^2$ estimates in meta-analysis.

It has been proposed that $I^2$ should be reported with their associated 95% CIs. By construct, the conventional frequentist CI represents the spectrum of results which would include the true underlying value in a particular proportion (typically 95%) if the experiment were independently repeated many times. In meta-analysis, we can conceptually think of an 'experiment' as a set of trials 'sampled' randomly from a universe of all possible trials. However, as outlined in the above sections, the patterns with which different types of trials are included in a meta-analysis over time are typically not random. For example, small trials are likely to precede larger trials. Thus, the statistical assumptions on which confidence intervals are based may not hold in many meta-analyses. For this reason, it is important to explore, empirically, how 95% confidence intervals perform as more trials are accumulated over time.

## Material

In a previous empirical study, we extracted data from 920 binary primary outcome meta-analyses in Cochrane systematic reviews.[22] We defined primary outcomes as one of the first three

outcomes in the first comparison group. The data set only included meta-analyses that pooled results across all trials; meta-analyses reporting only sub-totals were excluded. For this current study, we used the same population of meta-analyses and selected the subset of meta-analyses that met the following eligibility criteria:

- The total number of included trials surpassed 30. We employed this eligibility criterion because the number of trials is an important measure of the reliability of estimates of variation between trial results (i.e., $I^2$). Since we accept the final cumulative $I^2$ as representing a good approximation of the 'truth', it was important that the number of trials was large enough to make it likely that the final $I^2$ had converged and was stable.
- The total number of patients surpassed a required information size (i.e., required meta-analysis sample size) based on $\alpha$=5% and $\beta$=20% (i.e., 80% power). The required information size used for each meta-analysis was powered to detect a 25% relative risk reduction assuming a control group risk equal to the median of all trials. Calculation of a required information size requires an estimation of heterogeneity. For the purpose of estimating a reasonable required information size (and allowing confidence that our final effect estimate is reliable), we chose to assume a 50% degree of heterogeneity for these calculations.[23] $I^2$ is a function of Cochran's $Q$ and Cochran's $Q$ is a function of the sum of squared differences between each trial effect estimate and the meta-analysed effect estimate. Thus, if the meta-analysed effect estimate cannot be considered reliable, $I^2$ may not be reliable either.
- The disease of interest was a common disease. We employed this criterion because most interventions for common diseases yield intervention effects close to a 25% relative risk

reduction or smaller, thus giving credence to our considerations for the required information size (above criterion).

From the pool of 920 meta-analyses, 18 meta-analyses were originally eligible for our analysis, and after further considerations 16 studies were included. Post hoc, we elected to exclude two meta-analyses. These two meta-analyses each included two significantly different subgroups where all or the majority of trials in the second subgroup had been conducted after the trials in the first subgroup. We therefore did not find it appropriate to assess convergence of $I^2$ in this meta-analysis. Table 1 presents the characteristics of the 16 included meta-analyses.

## Analysis

For each of the 16 meta-analyses we calculated and plotted the cumulative $I^2$ estimate and associated 95% CI after each year of publication. We accepted the final $I^2$ estimate (i.e., the $I^2$ estimated based on the meta-analysis including all trials) as representing a good approximation of the 'truth'. First, we assessed the variation of $I^2$ estimates over time by calculating the difference between the maximum and minimum observed $I^2$ estimate over time in each meta-analysis. We refer to this difference as the *fluctuation span of* $I^2$. Second, we assessed how many events and trials were required for the cumulative $I^2$ estimate to become stable. We defined the considered $I^2$ estimates moderately and highly stable at the points where the cumulative $I^2$ estimate came within a +/-20% and a +/-10% absolute distance of the final cumulative $I^2$ estimate and stayed within this distance. Third, we recorded the cumulative number of trials and events where the 95% CIs temporarily did include the final $I^2$ estimate. At these time points, we assessed how far the closest CI limit was to the final $I^2$ estimate. That is, if the final $I^2$ estimate

was above the temporary 95% CI, we calculated the distance between the upper CI limit and the final $I^2$ estimate, and vice versa if the final $I^2$ estimate was below the 95% CI.

## Results

Columns 2-4 in table 2 present the minimum, the maximum, and the fluctuation span of $I^2$ values observed over time in each of the 16 included meta-analyses. The median, minimum and maximum fluctuation spans were 47.5%, 15%, and 81%. Ten of the 16 meta-analyses (62.5%) had a fluctuation span larger than 40%. Columns 5-8 in table 2 present the number of trials and events required for the cumulative $I^2$ estimate to become moderately and highly stable. In 3 of the 16 meta-analyses (meta-analyses 14-16) the cumulative $I^2$ estimates were moderately stable throughout the entire meta-analysis. For the remaining 13 meta-analyses, the median (minimum to maximum) number of trials and events required to become moderately stable was 11 (5 to 25) and 467 (138 to 1894) respectively. The median (minimum to maximum) number of trials and events required to become highly stable was 20 (10 to 37) and 958 (257 to 2766). Further, graphical inspection revealed that, except for one meta-analysis (meta-analysis 10), no major fluctuations occurred after the first point where the cumulative meta-analysis included at least 500 events and 15 trials.

In 3 of the 16 meta-analyses (meta-analyses 7, 9 and 14), the 95% CIs temporarily did not include the final $I^2$ estimate (see figures 1-4). In meta-analysis 7, the 95% CI at the second publication year was 0-69% and the final $I^2$ estimate was 77%. The cumulative number of events and trials at this point was 349 and 5. In meta-analysis 9, the 95% CI at the fourth publication year was 57-88% and the final $I^2$ estimate was 54%. The cumulative number of events and trials

at this point was 177 and 5. In meta-analysis 14, the 95% CI at the third and fourth publication year was 77-94% and 75-93% and the final $I^2$ estimate was 74%. The cumulative number of events and trials was 349 and 7 at the third year of publication and 407 and 8 at the fourth year of publication.

## Discussion

In summary, our findings suggest $I^2$ estimates are likely to incur considerable fluctuations when a meta-analysis includes less than roughly 500 events or 15 trials, and that 95% CIs for the $I^2$ estimate provide good coverage over time. All instances where the 95% CI temporarily did not include the final $I^2$ estimate occurred in cumulative meta-analyses with less than 500 events and 10 trials. However, even in the rare cases where the 95% CIs did not include the final $I^2$ estimate, it is unlikely that inferences about the degree of heterogeneity based on the temporary 95% CIs would have differed from inferences based on the final $I^2$ estimate.

Our study offers several strengths. First, it represents the first empirical evaluation of the evolution of $I^2$ estimates and their associated 95% confidence intervals over time. Second, our results provide novel insights on the use of one of the most important inferential measures, $I^2$, in meta-analytic practice. Third, we selected meta-analyses including a sufficiently large number of trials and patients to help ensure a sufficiently reliable sample.

Our study has a number of limitations. We only evaluated $I^2$ estimates and their associated 95% CIs after each year of publication. Since all of the included meta-analyses included more than 1 trial for some of the years, it is possible that the $I^2$ estimates in some of the 16 meta-analyses may

have become stable with a smaller number of events and trials than indicated in table 2. Some of the number of events and trials required to reach convergence which we present in table 2 may be therefore overestimates. Only 16 meta-analyses were eligible, having covered a limited spectrum of medical areas. Our findings may therefore not be generalizable to meta-analyses that bear little resemblance to the meta-analyses included in this study. Similarly, we also did not examine meta-analyses published in paper journals. A number of differences between Cochrane meta-analyses and journal based meta-analyses have been documented (e.g., meta-analyses published in paper journals are more likely to present statistically significant findings).[15,16] One could therefore speculate that fluctuations in $I^2$ estimates may differ between Cochrane and paper journal meta-analyses. In the second section of this paper (statistical framework and theoretical considerations), we explained that $I^2$ estimates may fluctuate due to lack of power, time-dependent bias, and evolving trial weights and precisions. We did not perform an in-depth assessment of the degrees to which each of these factors caused $I^2$ estimates to fluctuate in the 16 meta-analyses. We believe a simulation study would be more appropriate to explore this issue. Finally, we did not examine if any of the review authors took any precautions about uncertainty associated with $I^2$ estimates (especially in early versions of the systematic reviews where the meta-analysis included less than 500 events and 15 trials). However, given the paucity of methodological literature on the $I^2$ measure just five years ago, it is likely that most Cochrane review authors would have been unaware of the issues related to uncertainty associated with estimating $I^2$.

The median number of trials in a meta-analysis is 7 in Cochrane reviews and 12 in systematic reviews published in paper journals.[15,16] With clinical trial sample sizes typically being smaller

than 100,[17,18] it is likely that most published meta-analyses will incur considerable fluctuations (i.e., meta-analyses with less than 500 events and 15 trials). Hence, there is a need for presenting the $I^2$ estimate with its associated 95% CI. Unreliable $I^2$ estimates have potential negative implications for the assessment of reliability of intervention effect estimates. Recent literature as well as the GRADE initiative have promoted the need for assessing intervention effects in relation to the strength of evidence.[23-27] One of the factors when considering the overall quality of evidence is the precision of the pooled estimate of effect, which is achieved, in part, through considering the required (or optimal) information size.[23-27] However, to carry out such assessments reliably it is necessary to have a good idea of the expected degree of heterogeneity in the meta-analysis, and if this is not possible, one should at least carry out sensitivity assessments based on a plausible spectrum of degrees of heterogeneity. Uninformed use of the current $I^2$ estimate does not provide a solid basis for such assessments, but interpretation of the $I^2$ estimate in relation to the cumulative amount of evidence and the associated 95% CI does.

Previous studies have already identified limitations associated with the $I^2$ measure as well as the uncertainty associated with $I^2$ estimates.[7,8,11,12] Our study adds to the previous literature by introducing temporality. However, as pointed out above, our findings do have limitations and need confirmation in simulation studies and perhaps other empirical studies. Example papers, which put statistical inferences about the degree of heterogeneity in a clinical context, are also required. The latter may be realized if confidence intervals became an integral part of the widely used systematic review software Review Manager as well as other meta-analysis software packages.[28]

In conclusion, $I^2$ estimates are likely to fluctuate considerably in meta-analyses with less than roughly 500 events and 15 trials. Confidence intervals for $I^2$ estimates provide good coverage as evidence accumulates, and are thus valuable for reflecting the uncertainty associated with estimating $I^2$. It is our hope that the next updates of systematic review and meta-analysis software packages, such as Review Manager, will include confidence intervals for the $I^2$ estimate.

## Appendix

Let $Q$ be Cochran's homogeneity test statistic and let $k$ be the number of trials included in a meta-analysis. The calculation of confidence intervals for $I^2$ using the 'test based' methods relies on a transformation of the $I^2$ statistic, $H^2=1/(1- I^2)=Q/(k-1)$. The test based method utilizes that the statistic $Z=\sqrt{(2Q)}-\sqrt{(2k-3)}$ approximately follows a standard normal distribution. Similarly, if we take the natural logarithm of $Q$ we remove much of the skewness of the underlying distribution of $Q$. Because the expectation of $Q$ is $k-1$, $(ln(Q)-ln(k-1))/(SE(ln(Q)))$ can be assumed to approximately follow a standard normal distribution. Equating Z with this expression and isolating for $SE(ln(Q))$ we get

$$SE(\ln(Q)) = \frac{\ln(Q) - \ln(k-1)}{\sqrt{2Q} - \sqrt{2k-3}}$$

And since $Q=(k-1)H^2$ and $k$ is a constant we have $SE(ln(Q))=1/2\ SE(ln(H))$, and hence

$$SE(\ln(H)) = \frac{1}{2}\frac{\ln(Q) - \ln(k-1)}{\sqrt{2Q} - \sqrt{2k-3}} \qquad (1)$$

One problem with this approach is that the standard error approaches zero as $H$ approaches 1. Small values of $H$ indicate homogeneity of trial results in which case $Q$ is $\chi^2$ distributed with $k-1$ degrees of freedom. In this case we can take an approximate variance of $ln(Q/(k-1))=2*ln(H)$ and arrive at

$$SE(\ln(H)) = \sqrt{\frac{1}{2(k-2)}\left(1 - \frac{1}{3(k-1)^2}\right)} \qquad (2)$$

Higgins et all showed via simulation that formula (1) should be used when Q>k and formula (2) should be used when $Q \leq k$. Having estimated the standard error of $ln(H)$ one can assume approximate normality and derive approximate 95% confidence intervals for $H$: $exp(ln(H) \pm 1.965 * SE(ln(H)))$. One can subsequently square the resulting CI limits and transform each of them back to a percentage of heterogeneity, $I^2$.

## References


(1) Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Statistics in Medicine* 2000; 19:1707-1728.

(2) Higgins JP, Thompson S. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 2002; 21:1539-1558.

(3) Higgins JP, Thompson S, Deeks J, Altman DG. Measuring inconsistency in meta-analyses. *British Medical Journal* 2003; 327:557-560.

(4) Higgins JP, Green S. Cochrane Handbook for systematic reviews of interventions, version 5.0.0. John Wiley & Sons; 2009.

(5) Lau J, Ioaniddis JP, Schmid CH. Summing up evidence: one answer is not always enough. *Lancet* 1998; 351:123-127.

(6) Ioaniddis JP. Interpretation of tests of heterogeneity and bias in meta-analysis. *Journal of Evaluation of Clinical Practice* 2008; 14:951-957.

(7) Huedo-Medina T, Sánchez-Meca J, Marín-Martínez F. Assessing heterogeneity in meta-analysis: Q Statistic or I(2) Index? *Psychological Methods* 2006; 11(2):193-206.

(8) Mittlböck M, Heinzl H. A simulation study comparing properties of heterogeneity measures in meta-analysis. *Statistics in Medicine* 2010; 25:4321-4333.

(9) Jackson D. The implications of publication bias for meta-analysis' other parameter. *Statistics in Medicine* 2006; 25(17):2911-2921.

(10) Jackson D. Assessing the implications of publication bias for two popular estimates of between-study variance in meta-analyses. *Biometrics* 2007; 63(1):187-193.

(11) Rücker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I(2) in assessing heterogeneity may mislead. *BMC Medical Research Methodolgy* 2008;(8):79.

(12) Ioannidis JP, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analysis. *BMJ* 2007; 335:914-916.

(13) Guyatt G, Wyer P, Ioaniddis.JP. When to Believe a Subgroup Analysis. Users' Guide to the Medical Literature: A manual for Evidence-Based Clinical Practice. McGraw-Hill; 2008.

(14) Cochran WG. The combination of estimates from different experiments. *Biometrics* 1954; 10:101-129.

(15) Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. *PLoS Medicine* 2010; 4(3):e78.

(16) Tricco AC, Tetzlaff J, Pham B, Brehaut J , Moher D. Non-Cochrane vs. Cochrane reviews were twice as likely to have positive conclusion statements: cross-sectional study. *Journal of Clinical Epidemiology* 2009; 62(4):380-386.

(17) Chan AW, Altman DG. Epidemiology and reporting of randomized clinical trials published in PubMed journals. *Lancet* 2005; 365(9465):1159-1162.

(18) Gluud C. The culture of designing hepato-biliary randomised clincial trials. *Journal of Hepatology* 2006; 44:607-615.

(19) Dickersin K. The existence of publication bias and risk factors of its occurence. *Journal of American Medical Association* 1990; 263:1385-1389.

(20) Dwan K, Alman D, Arnaiz J, Bloom J, Chan AW, Cronin E et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS Medicine* 2008; 3:e3081.

(21) Ioaniddis JP. Effect of Statistical Significance of Results on the Time to Completion and Publication of Randomized Efficacy Trials. *Journal of American Medical Association* 1998; 279(4):281-286.

(22) Thorlund K, Wetterslev J, Awad T, Thabane L, Gluud G. Comparison of statistical inferences from the DerSimonian-Laird and alternative random-effects model meta-analyses - an empirical assessment of 920 primary outcome Cochrane meta-analyses. *Research Synthesis Methods* 2011; Submitted.

(23) Wetterslev J, Thorlund K, Brok J, Gluud C. Estimating required information size by quantifying diversity in a random-effects meta-analysis. *BMC Medical Research Methodology* 2009; 9(86).

(24) GRADE Working Group. Grading quality of evidence and strength of recommendations in clinical practice guidelines. Part 1 of 3. An overview of the GRADE approach and grading quality of evidence about interventions. *Allergy* 2009; 64:669-677.

(25) Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008; 336(7650):924-926.

(26) Thorlund K, Devereaux PJ, Wetterslev J, Guyatt G, Ioannidis JP, Thabane L et al. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? *International Journal of Epidemiology* 2009; 38:276-286.

(27) Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. *Journal of Clinical Epidemiology* 2008; 61:64-75.

(28) Review Manager (RevMan) [Computer program]. Version 5.0. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2008.

# Tables

| Disease/population | Outcome | Experimental intervention | Control Intervention | Period | Cumulative (final) statistics | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Trials | Events | Patients | $I^2$ |
| (1) Colon Cancer | Death | Adjuvant therapy for completely resected stage II cancer | No adjuvant therapy | 1987-2007 | 44 | 3402 | 17805 | 0% |
| (2) Need for perioperative allogeneic blood transfusion | Exposure to allogeneic blood | Aprotinin | Blood transfusion & blood loss | 1987-2006 | 96 | 5348 | 10144 | 68% |
| (3) Bacterial infections in afebrile neutropenic patients following chemotherapy | Febrile patients/episodes | Antibiotic prophylactic drugs | Placebo/no intervention | 1973-2005 | 46 | 3201 | 6023 | 74% |
| (4) Fever following cesarean section | Fever | Antibiotic prophylaxis | Control | 1971-2001 | 45 | 1504 | 7180 | 49% |
| (5) Postoperative infection after appendectomy | Wound infection | Antibiotics | Placebo | 1986-1995 | 70 | 919 | 8812 | 26% |
| (6) Pre-eclampsia and its complications | Gestational hypertension | Antiplatelet agents | Placebo/No antiplatelet agents | 1985-2004 | 33 | 2080 | 20701 | 48% |
| (7) Need for perioperative allogeneic blood transfusion | Exposure to allogeneic blood | Cell salvage | Blood transfusion & blood loss | 1979-2003 | 46 | 1808 | 3857 | 77% |
| (8) Smokers | Smoking cessation at 6+ months follow-up | Nicotine replacement therapy (any type) | Placebo/No therapy control | 1979-2007 | 111 | 5962 | 43040 | 23% |
| (9) Smokers | Smoking cessation at longest follow-up | Nursing interventions | Control | 1987-2005 | 31 | 1977 | 15205 | 54% |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| (10) Colorectal cancer | Recurrence of cancer | Perioperative blood transfusion | No intervention | 1985-2001 | 36 | 4026 | 12127 | 59% |
| (11) Chronic hepatitis C | Sustained virological response | Ribavirin plus interferon | Interferon | 1995-2004 | 54 | 6126 | 8354 | 80% |
| (12) Rapid sequence induction intubation | Intubation condition | Rucoronium | Succinylcholine | 1992-2006 | 37 | 1948 | 2690 | 55% |
| (13) Non-small cell lung cancer in patients with advanced disease | Response to treatment | Double agent regimens | Single agent regimen | 1984-2003 | 33 | 1410 | 7175 | 53% |
| (14) Metastatic breast cancer | Response to treatment | Single agent | Combination chemotherapy | 1975-2003 | 38 | 2380 | 6184 | 75% |
| (15) Postoperative pain in adults | >50% pain relief over 4 to 6 hours | Single dose oral paracetamol | Placebo | 1975-2006 | 56 | 1969 | 5762 | 63% |
| (16) Pregnant women at labor term | Caesarean section | Vaginal prostaglandin (for induction) | Placebo/No treatment | 1979-1997 | 31 | 898 | 6243 | 0% |

**Table 1** Characteristics of the 16 included meta-analyses.

| Meta-analysis | Minimum $I^2$ value | Maximum $I^2$ value | Fluctuation span of $I^2$ | Number of trials required to become stabile | | Number of events required to become stabile | |
|---|---|---|---|---|---|---|---|
| | | | | Moderately (+/-20%) | Highly (+/-10%) | Moderately (+/-20%) | Highly (+/-10%) |
| (1) | 0% | 24% | 24% | 24 | 33 | 1297 | 2096 |
| (2) | 42% | 74% | 32% | 11 | 35 | 276 | 2766 |
| (3) | 46% | 74% | 28% | 5 | 19 | 149 | 924 |
| (4) | 2% | 54% | 52% | 9 | 28 | 287 | 992 |
| (5) | 0% | 51% | 51% | 25 | 34 | 335 | 453 |
| (6) | 0% | 51% | 51% | 16 | 16 | 562 | 562 |
| (7) | 24% | 78% | 54% | 12 | 12 | 597 | 597 |
| (8) | 0% | 48% | 48% | 11 | 33 | 610 | 1509 |
| (9) | 0% | 77% | 77% | 11 | 21 | 537 | 1393 |
| (10) | 24% | 69% | 45% | 18 | 18 | 1894 | 1894 |
| (11) | 0% | 81% | 81% | 6 | 12 | 138 | 1989 |
| (12) | 10% | 57% | 47% | 13 | 37 | 467 | 1948 |
| (13) | 0% | 63% | 63% | 8 | 18 | 199 | 580 |
| (14) | 65% | 88% | 23% | - | 10 | - | 475 |
| (15) | 51% | 66% | 15% | - | 24 | - | 879 |
| (16) | 0% | 18% | 18% | - | 13 | - | 257 |

**Table 2** Presents the fluctuation span of $I^2$ values and the number of trials and events required to become stabile.

# Figures

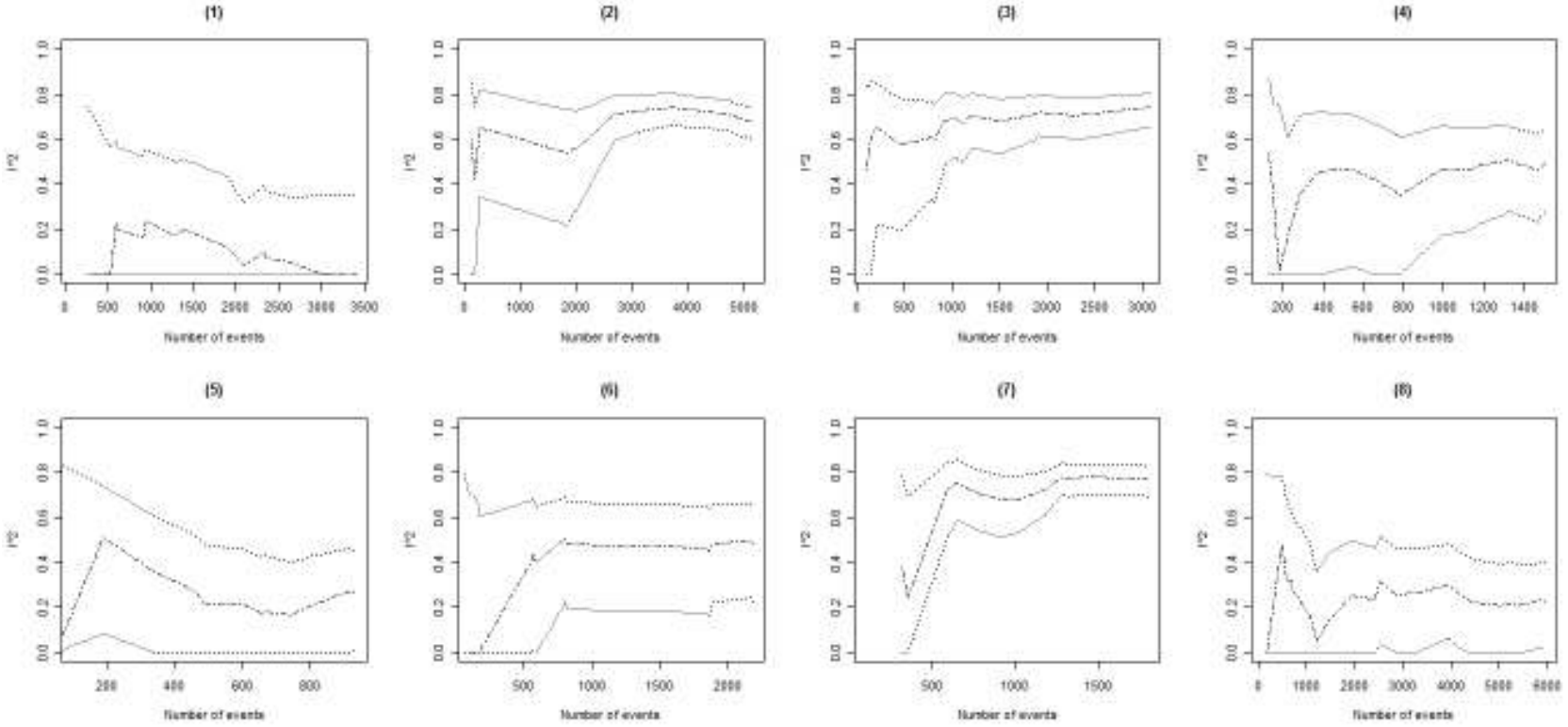



**Figure 1** Presents the evolution of the cumulative $I^2$ estimates and their associated 95% confidence intervals (CIs) over the accumulation of events in meta-analyses (1) to (8). The cumulative $I^2$ estimates are represented by the dot-dashed line (– · – · ), and their associated cumulative 95% CIs are represented by the dotted lines ( ·····).

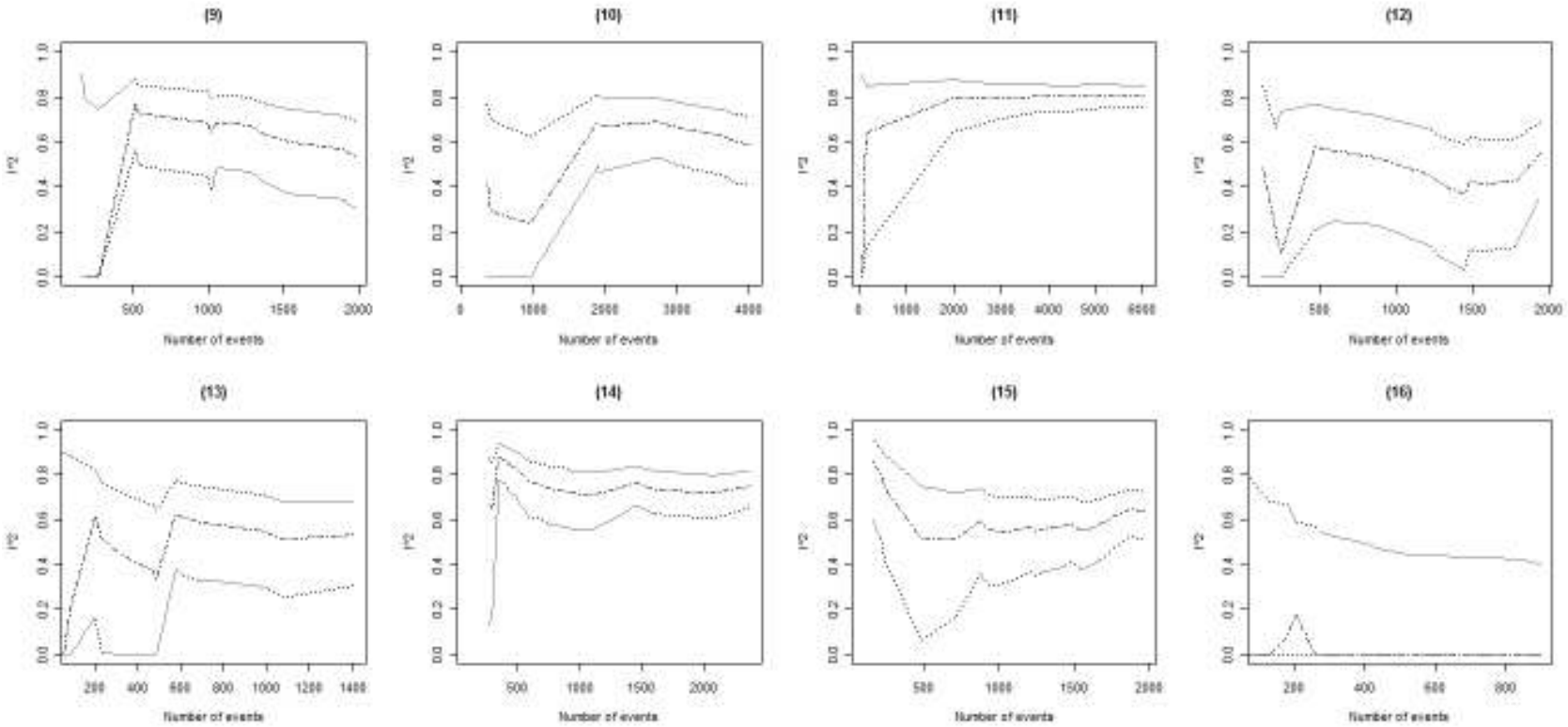



**Figure 2** Presents the evolution of the cumulative $I^2$ estimates and their associated 95% confidence intervals (CIs) over the accumulation of events in meta-analyses (9) to (16). The cumulative $I^2$ estimates are represented by the dot-dashed line (– · – · ), and their associated cumulative 95% CIs are represented by the dotted lines (·····).

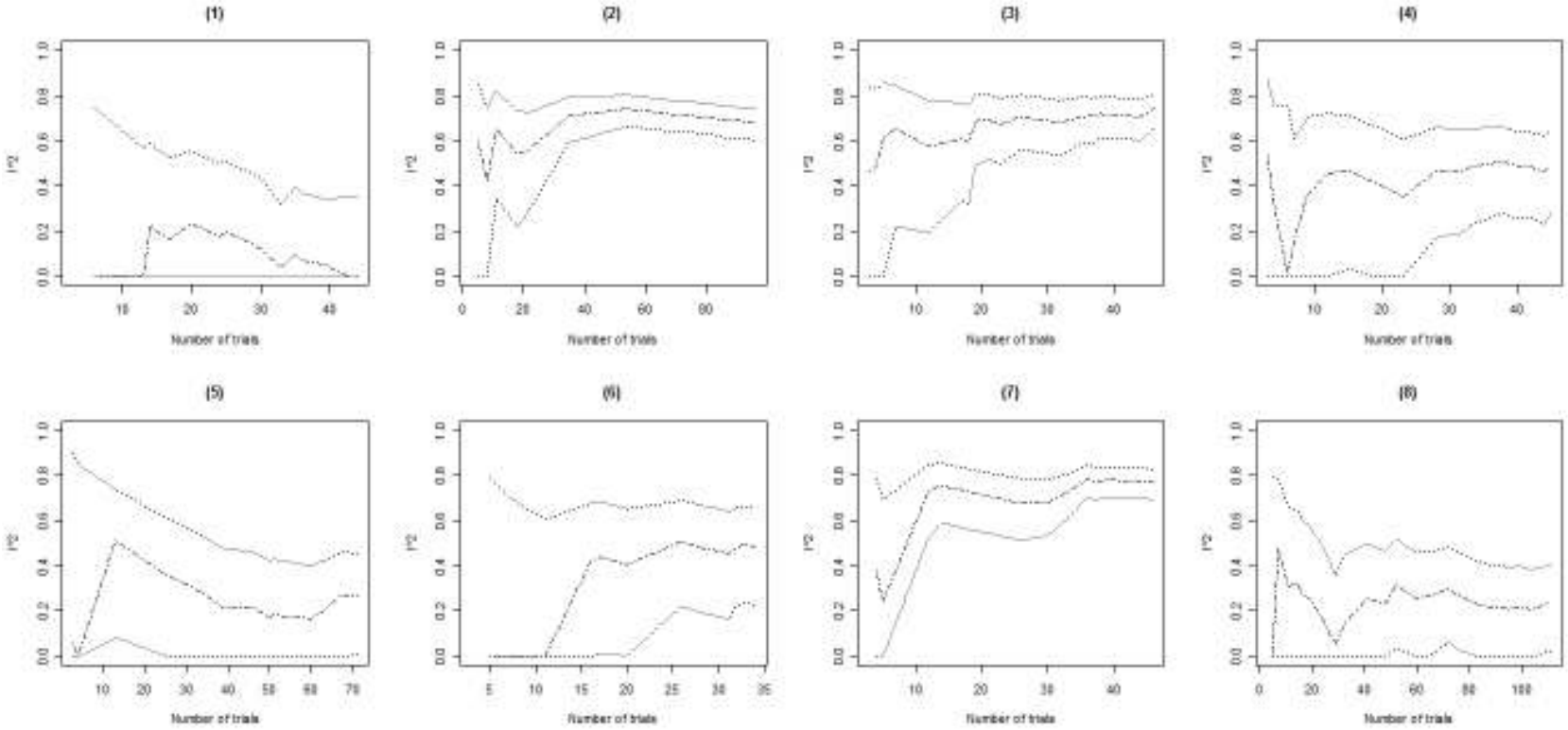



**Figure 3** Presents the evolution of the cumulative $I^2$ estimates and their associated 95% confidence intervals (CIs) over the accumulation of trials in meta-analyses (1) to (8). The cumulative $I^2$ estimates are represented by the dot-dashed line (– · – ·), and their associated cumulative 95% CIs are represented by the dotted lines (·····).

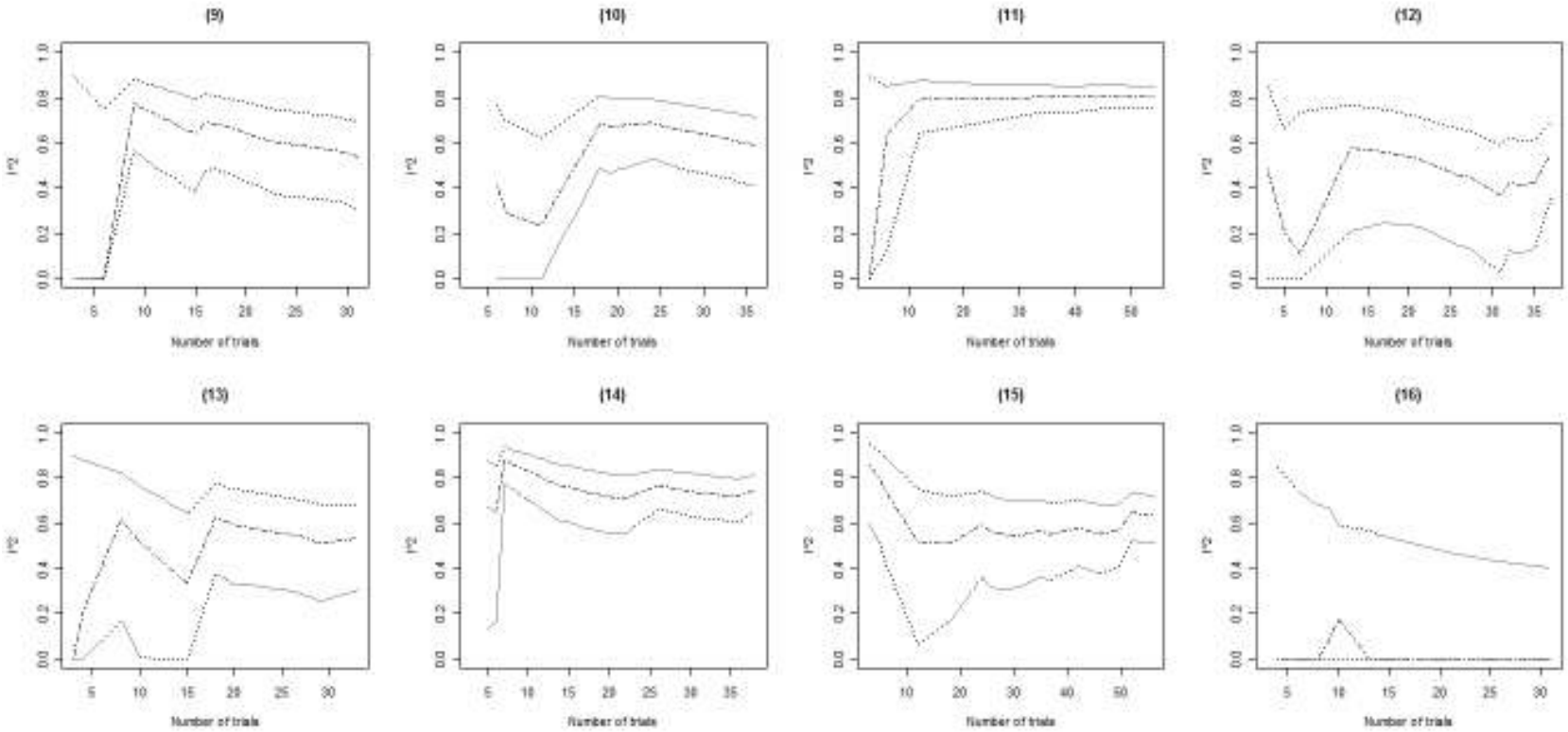



**Figure 4** Presents the evolution of the cumulative $I^2$ estimates and their associated 95% confidence intervals (CIs) over the accumulation of trials in meta-analyses (9) to (16). The cumulative $I^2$ estimates are represented by the dot-dashed line (– · – · ), and their associated cumulative 95% CIs are represented by the dotted lines (·····).

# Chapter 4: Comparison of statistical inferences from the DerSimonian-Laird and alternative random-effects model meta-analyses – an empirical assessment of 920 Cochrane primary outcome meta-analyses

Authors:

Kristian Thorlund

Jørn Wetterslev

Tahany Awad

Lehana Thabane

Christian Gluud

Word Count:

Summary: 190

Manuscript: 7033

## Summary

### Background

In random-effects model meta-analysis, the conventional DerSimonian-Laird (DL) estimator typically underestimates the between-trial variance. Alternative variance estimators have been proposed to address this bias.

### Objectives

To empirically compare statistical inferences from random-effects model meta-analyses based on the DL estimator and four alternative estimators, as well as distributional assumptions (normal and t-distribution) about the pooled intervention effect.

### Methods

We evaluated discrepancies of p-values, 95% confidence intervals in statistically significant meta-analyses, and the degree (percentage) of statistical heterogeneity (e.g., $I^2$) across 920 Cochrane primary outcome meta-analyses.

### Results

In total, 409 of the 920 meta-analyses were statistically significant with the DL meta-analysis and 511 were not. Compared to the DL estimator, the four alternative estimators yielded p-values and confidence intervals that could be interpreted as discordant in up to 11.6% or 6% of the included meta-analyses pending whether a normal- or a t-distribution of the intervention effect estimates were assumed. Large discrepancies were observed for the measures of degree of heterogeneity when comparing DL to each of the four alternative estimators.

### Conclusion

Estimating the degree (percentage) of heterogeneity based on less biased between-trial variance estimators seems preferable to current practice. Disclosing inferential sensitivity of p-values and confidence intervals may also be necessary when borderline significant results have substantial impact on the conclusion.

## Introduction

Meta-analysis combining the results of randomised clinical trials is considered the best available estimate of an intervention effect.[1] Trials in a meta-analysis are often heterogeneous and need to be analysed accordingly.[1-5] When the underlying sources of heterogeneity between trials are unidentifiable and therefore cannot be explained, a random-effects model may be adopted to obtain an average estimate of the intervention effects.[1-4] Currently, the vast majority of published systematic reviews including meta-analysis use the well-known random-effects model proposed by DerSimonian and Laird (DL), which entails estimating the between-trial (heterogeneity) variance with the DL estimator, and obtaining confidence intervals and p-values under the assumption that the meta-analysed intervention effect estimate follows a normal distribution.[2] Several simulation studies have demonstrated that the DL estimator is likely to underestimate the between-trial variance.[6-12] When this happens, the p-value for the meta-analysed intervention effect may become artificially small and the confidence interval (CI) artificially narrow. Review authors therefore risk falsely concluding that an experimental intervention is effective. Furthermore, underestimation of the between-trial variance - and thus underestimation of statistical heterogeneity - may inappropriately draw the authors' and readers' attention away from the need to explore heterogeneity. These issues raise concern as to whether or not the statistical inferences from DL random-effects model meta-analyses in systematic reviews are appropriate.

Several alternative between-trial variance estimators are available - some of which have been demonstrated to produce less downward biased estimates, and thus, more accurate p-values and confidence intervals.[6-11;13;14] Further, robust random-effects model variance estimation, which

assumes the meta-analysed intervention effect estimate follows a t-distribution, has been shown to yield reliable confidence intervals (i.e., consistently good coverage) that are less sensitive to the choice of a specific between-trial variance estimator.[8;12;15-17] Unbiased variance and heterogeneity estimation is especially important in meta-analyses of primary outcomes because they typically shape the conclusions of systematic reviews. The superiority of these alternative variance estimators has only been demonstrated through simulation studies. In practice, it is unclear how often use of these alternative variance estimators will change the estimated p-value, the confidence interval, and the magnitude of the heterogeneity to the extent where the statistical inferences change.

## Objectives

To inform the above mentioned issues we performed a large scale empirical evaluation of primary outcome meta-analyses (see definition in data extraction section) from systematic reviews of clinical trials published in The Cochrane Library. We furthermore provide two illustrative examples where the use of alternative variance estimators could potentially have impacted the conclusions drawn in the original systematic reviews.

## Methods

In this section we first outline the statistical framework for the random-effects model. We then provide a description of the DL estimator followed by a description of four alternative estimators. The alternative estimators were selected among approximately 15 estimators we came across in the literature.[2;6;8-11;13-15;18] The alternative estimators were selected based on the fact that previous studies had shown them to be less biased than the DL estimator or that they

had received attention in recent simulation studies. We limited our analyses to the four estimators to retain simplicity in the presentation of results.[2;8-10;12;13]

Lastly, we describe the measures we employ to assess the extent and frequency with which the use of alternative between-trial variance estimators in meta-analysis may cause important changes in the statistical inferences.

**The conventional random-effects meta-analysis model**

Assume we have $k$ independent trials. Let $Y_i$ denote the estimate of the effect from $i$th trial. Let $\mu_i$ be the true intervention effect of the $i$th trial, and let $\sigma_i^2$ denote the variance of $\mu_i$. The trial specific intervention effects are assumed to vary across trials, with an underlying true effect $\mu$, and a between-trial variance $\tau^2$. In the random-effects model the observed effect measure, $Y_i$, is then assumed to satisfy the distributional relationship $Y_i \sim N(\mu, \sigma_i^2 + \tau^2)$, and the trial weights, $w_i^*$, are set at the inverse of the trial variances $w_i^* = 1/(\sigma_i^2 + \tau^2)$. In practice neither $\sigma_1^2, \ldots, \sigma_k^2$, nor $\tau^2$ are known. The within-trial variances are typically estimated using the trial sampling variances $\hat{\sigma}_1^2, \ldots, \hat{\sigma}_k^2$ and the between-trial variance $\hat{\tau}^2$ is obtained using a specific estimator (e.g., the DL estimator). The pooled intervention effect is obtained as a weighted pooled estimate of the observed intervention effects in the included trials

$$\hat{\mu}_w = \frac{\sum_{i=1}^{k} \hat{w}_i^* \cdot Y_i}{\sum_{i=1}^{k} \hat{w}_i^*} \qquad (1)$$

its standard error (SE) is estimated as $SE\left(\hat{\mu}_w\right) = \sqrt{1/\sum_i \hat{w}_i^*}$ , and the two-sided (1-α/2)% CI is given by $\hat{\mu}_w \pm z_{1-\alpha/2} \cdot SE(\hat{\mu}_w)$, where $z_{1-\alpha/2}$ is the (1-α/2)th percentile from the standard normal distribution. Thus, the conventional 95% CI is given by $\hat{\mu}_w \pm 1.96 \cdot SE(\hat{\mu}_w)$.

**Weighted variance estimation assuming a t-distribution**

A promising alternative which has not yet been used widely in practice consists of assuming that the pooled treatment effect follows a t-distribution and subsequently calculating a 'weighted extension' of the general formula for the variance of the pooled treatment effect.[16;19]

$$Var_w\left(\hat{\mu}_w\right) = \frac{\sum_i \hat{w}_i^* (Y_i - \hat{\mu}_w)^2}{(k-1)\sum_i \hat{w}_i^*}$$

And as always we have $SE_w\left(\hat{\mu}_w\right) = \sqrt{Var_w\left(\hat{\mu}_w\right)}$ . Hartung showed that the statistic $t = \left(\hat{\mu}_w - \mu\right) / SE_w\left(\hat{\mu}_w\right)$ follows a t-distribution with *k-1* degrees of freedom. Thus, two-sided (1-α/2)% CI based on the weighted variance is given by $\hat{\mu}_w \pm t_{k-1,1-\alpha/2} \cdot SE(\hat{\mu}_w)$, where $t_{k-1,1-\alpha/2}$ is the (1-α/2)th percentile from a t-distribution with k-1 degrees of freedom.

This approach has demonstrated robustness in simulation studies, in which, the confidence intervals have generally provided actual coverage close to the nominal coverage,[8;9;15;17] and test statistics have generally exhibited good control of the desired type I error rate.[11;20;21] Further, simulation studies have also suggested that results under the 'weighted variance' approach are less affected by the choice of between-trial variance estimator, than the conventional random-

effects approach which assumes the meta-analysed intervention effect estimates follow a normal distribution (see above).[8;9;15;17]

**Random-effects model between-trial variance estimators**

For this study we considered the DL estimator and the four alternative between-trial variance estimators described below.[2;8-10;12;13]

***The DerSimonian-Laird (DL) between-trial variance estimator***

In the random-effects model approach proposed by DL, the between-trial variance is estimated using a method of moments based estimator.[2] Let $\hat{w}_i = 1/\hat{\sigma}_i^2$ denote the estimated weights under a fixed-effect model, and let $\hat{\mu}_{FE} = \left(\sum_i \hat{w}_i \cdot Y_i\right) / \left(\sum_i \hat{w}_i\right)$ be the weighted mean effect size under the fixed-effect model. Cochran's homogeneity test statistic, $Q = \sum_i \hat{w}_i (Y_i - \hat{\mu}_{FE})^2$, is used as the basis of the DL estimator, $\hat{\tau}^2_{DL}$, since its 1st moment takes the form $E(Q) = (k-1) + \hat{\tau}^2_{DL}\left(S_1 - (S_2 / S_1)\right.$, where $S_1 = \sum_i \hat{w}_i$ and $S_2 = \sum_i \hat{w}_i^2$. Isolating for $\hat{\tau}^2_{DL}$ then yields the expression for the moment estimator of the between-trial variance

$$\hat{\tau}^2_{DL} = \max\left(0, \frac{Q-(k-1)}{S_1 - S_2/S_1}\right) \qquad (2)$$

Simulation studies have shown that the DL estimator works well for meta-analyses that are subject to ignorable or little heterogeneity, but tends to underestimate the between-trial variance in meta-analyses that are subject to moderate or substantial heterogeneity.[6-8;10;15] This

underestimation may typically result in poor control of the type I error and poor coverage of the associated random-effects confidence intervals.[6-8;11;15] For example, normal distribution based 95% confidence intervals based on the DL estimator will often provide coverage anywhere between 80% and 93% depending on the meta-analysis scenario.[6-8;15] One comprehensive simulation study measured the mean square error (MSE) of seven estimators and found that except for meta-analyses with extreme heterogeneity, the DL estimator is less variable than the six other estimators.[10] The difference in variability between the DL estimator and other estimators is inversely correlated with the number of trials in a meta-analysis.[10] The authors of the study nevertheless commented that the reduced variability in meta-analysis including only a few trials may be explained by the facts that $Q$ is underpowered when the number of trials is small, in which case $\hat{\tau}^2_{DL}$ is truncated to 0 when $Q<k-1$.[10]

***The Hartung and Makambi (HM) estimator***

The Hartung and Makambi (HM) estimator is a modification of the DerSimonian-Laird estimator which is always positive (i.e., does not need to be truncated to zero when the estimate is smaller than zero).[12] It is given by

$$\tau^2_{HM} = \frac{Q^2}{\left(2\cdot(k-1)+Q\right)\cdot\left(S_1-(S_2/S_1)\right)} \qquad (3)$$

Where $Q$, $S_1$, and $S_2$ are given as above. One simulation study showed that the HM estimator exhibits better control over the type I error than the DL estimator.[11] However, in some scenarios the HM estimator still did not provide type I error close to the desired level. Another simulation

study showed that the HM estimator yields better or similar confidence interval coverage compared to the DerSimonian-Laird estimator in continuous data meta-analysis scenarios.[8]

***The restricted maximum likelihood (REML) estimator***

Restricted maximum likelihood (REML) estimation is a generally well-known estimation technique in the statistical literature. The between-trial variance REML estimator in meta-analysis is not widely used in practice, but tends to be included in simulation studies exploring aspects of heterogeneity estimation in meta-analysis, and can also be used for between-trial variance estimation in meta-regression. The REML between-trial variance estimate, $\hat{\tau}^2_{REML}$, is obtained through a double-iterative process. The first iteration involves (iterative) estimation of the maximum likelihood (ML) estimator of the between-trial variance, $\hat{\tau}^2_{ML}$, which is given by

$$\hat{\tau}^2_{ML} = \frac{\sum_{i=1}^{k} (\hat{w}_i^*)^2 \left( \left( Y_i - \hat{\mu}_{ML} \right)^2 - \hat{\sigma}_i^2 \right)}{\sum_{i=1}^{k} (\hat{w}_i^*)^2} \qquad (4)$$

where $\hat{w}_i = 1/(\hat{\sigma}_i^2 + \hat{\tau}^2)$, $\hat{\tau}^2$ is any estimate of the between-trial variance, and $\hat{\mu}_{ML}$ is the weighted pooled estimate of the intervention effect obtained as in equation (1) but with weights $\hat{w}_i$ instead of $\hat{w}_i^*$.[8;10] The ML estimator is obtained by iterating over $\hat{\tau}^2_{ML}$ and $\hat{\mu}_{ML}$, until $\hat{\tau}^2_{ML}$ has converged. If we let $\hat{\tau}^2_{ML(i)}$ and $\hat{\tau}^2_{ML(i-1)}$ denote the current and the preceding iterations' estimates, one typical convergence criterion could be that $|\hat{\tau}^2_{ML(i)} - \hat{\tau}^2_{ML(i-1)}| / (1 + \hat{\tau}^2_{ML(i-1)}) < 0.0001$.[8;10] The initial value of the between-trial variance, $\hat{\tau}^2_0$, can be estimated with any other non-iterative

estimator or chosen as any plausible value respective to the scale on which trial results are being pooled (e.g., the log odds ratio scale). The second iteration round uses $\hat{\tau}^2_{ML}$ and $\hat{\mu}_{ML}$ as initial estimates for $\hat{\tau}^2_{REML}$ and $\hat{\mu}_{REML}$ to estimate $\hat{\tau}^2_{REML}$

$$\hat{\tau}^2_{REML} = \frac{\sum_{i=1}^{k} (\hat{w}_i^*)^2 \left( \left( Y_i - \hat{\mu}_{REML} \right)^2 - \hat{\sigma}_i^2 \right)}{\sum_{i=1}^{k} (\hat{w}_i^*)^2} - \frac{1}{\sum_{i=1}^{k} \hat{w}_i^*} \qquad (5)$$

Simulation studies have found that the REML estimator is typically less downwardly biased than the DL estimator.[6;10] One should, however, note that these simulation studies also suggest the REML estimator is sub-optimal compared to other proposed estimators. [6;10] One simulation study showed that the REML estimator was generally equally or slightly more variable (measured as associated MSE) than the DL estimator.[10]

***The Hedges (HE) estimator***

The Hedges (HE) estimator is an analogue to the variance components estimator in a random-effects analysis of variance (ANOVA).[13] The HE estimator is a simple unweighted variance estimator given by

$$\hat{\tau}^2_{HE} = \frac{\sum_{i=1}^{k} \left( Y_i - \hat{\mu}_{uw} \right)^2}{k-1} - \frac{\sum_{i=1}^{k} \hat{\sigma}_i^2}{k} \qquad (6)$$

where $\hat{\mu}_{uw}$ is the unweighted mean of the observed trial effect estimates, $Y_i$. Simulation studies have shown that under the assumption that the pooled intervention effect follows a normal distribution, the HE estimator works well for meta-analyses that are subject to substantial heterogeneity, but tends to overestimate the between-trial variance in meta-analyses that are subject to ignorable or little heterogeneity.[8;10] One simulation study showed that the HE estimator was generally considerably more variable than the DL estimator.[10] For example, in simulated meta-analyses with fewer than 20 trials where the heterogeneity was not extreme, the MSE of the HE estimator was 25% to 100% larger than the MSE of the DL estimator across various scenarios.[10]

***The Sidik and Jonkman (SJ) estimator***

The Sidik and Jonkman (SJ) estimator is based on the usual statistical method for estimating the model error variance in a weighted linear model.[9] The SJ estimator is given by

$$\hat{\tau}^2_{SJ} = \frac{\sum_{i=1}^{k} \hat{v}_i \left(Y_i - \hat{\mu}_{w0}\right)^2}{k-1} \qquad (7)$$

where $\hat{v}_i = (\hat{\sigma}_i^2 / \hat{\tau}_0^2) + 1$, $\hat{\tau}_0^2$ is an initial estimate of the between-trial variance, typically the HE estimator, and $\hat{\mu}_{w0}$ is the weighted random-effects pooled estimate using $\tau_0^2$ as the estimate for the between-trial variance. When $\hat{\tau}_0^2$ is truncated to zero, one may instead set $\hat{\tau}_0^2$ to 0.01. Simulation studies have shown that the SJ estimator is generally the least downwardly biased among estimators, and authors have previously recommended this estimator as the preferred choice for estimating the between-trial variance.[8-10] The SJ estimator particularly works well for

meta-analyses that are subject to mild or moderate heterogeneity, but may yield slight underestimates for meta-analyses that are subject to substantial heterogeneity.[10] One simulation study, however, showed that the SJ estimator was generally more variable than the DL estimator.[10] For example, in simulated meta-analyses with fewer than 20 trials where the heterogeneity was not extreme, the MSE of the SJ estimator was 20% to 50% larger than the MSE of the DL estimator across various scenarios.[10]

**Measuring frequency and magnitude of changes in statistical inferences**

Between-trial variance estimates from different estimators are seldom equal. Random-effects model meta-analyses based on different estimators will therefore typically not yield identical p-values, confidence intervals, percentage estimates of heterogeneity, and other summary statistics. However, differences between summary statistics are only important if they are large enough to alter the inferences drawn about the investigated intervention effect. Below we describe the measures we employ to assess differences between the p-values, confidence intervals, and estimated degrees of heterogeneity resulting from the considered random-effects model approaches.

***P-values and agreement in statistical significance***

P-values are commonly assessed on the basis of whether they are smaller than some given threshold – that is, whether the results can be considered statistically significant or not. Conventionally meta-analyses use a threshold of 0.05. We therefore assessed agreement between p-values based on this threshold. That is, if the p-values from two random-effects meta-analyses based on different between-trial variance estimators lie on the same side of 0.05, we considered

the two estimators to agree on statistical significance. If the p-values lie on each side of 0.05, we considered the two estimators to disagree on statistical significance.

Disagreement on statistical significance, as defined above, may however not be sufficiently pronounced to alter the inferences about the overall treatment effect (and thus, possibly the conclusion of the systematic review). For example, if one p-value based on one between-trial variance estimator yields a borderline statistically significant result (e.g., p=0.03) and another p-value based on another between-trial variance estimator yields a borderline 'not statistically significant' result (e.g., p=0.07), the disagreement on statistical significance may not be sufficiently pronounced to alter the statistical inferences about the overall treatment effect, and thus, to impact the conclusion of the systematic review. Conversely, in a meta-analytic scenario where the smallest of the two p-values yields a strongly significant result (e.g., p=0.002) or where the largest p-value yields clear absence of statistical significance (e.g., p=0.36), using one p-value instead of the other is more likely to alter the inferences about the overall treatment effect and thus the conclusion of the systematic review. There are many potential ways of assessing how pronounced disagreement on statistical significance is. For this paper, we chose, as post-hoc analyses, to group p-values into four categories: '$p>0.10$', '$p\leq0.10$ and $p>0.05$', '$p\leq0.05$ and $p>0.01$' and '$p\leq0.01$' – each representing what can roughly be interpreted as 'weak', 'borderline', 'moderate', and 'strong' statistical significance. If two p-values (from two different estimators) fall into categories next to each other, this may not be inferentially problematic, whereas two p-values that are two or three categories apart may have considerable potential to affect the meta-analytic inferences, and thus, the conclusion of the systematic review.

***Agreement between confidence intervals in statistically significant meta-analyses***

If two random-effects model meta-analyses based on two different between-trial variance estimators are both statistically significant, but the widths of the two resulting CIs differ considerably, those two meta-analyses are likely to render different inferences about the investigated intervention. For example, if one between-trial variance estimator yields a borderline significant meta-analysis (e.g., the estimated relative risk of a clinical important event is less than 1.00 and the upper limit of the 95% CI is 0.99) and another between-trial variance estimator yields a 95% CI that precludes all clinically irrelevant intervention effects (e.g., the upper limit of the 95% CI is 0.90), the two estimators will typically render different inferences about the effect of the investigated intervention – or at least about the strength of evidence supporting the observed intervention effect. Investigators are likely to perceive the importance of differences between CIs differently. For this study, we arbitrarily defined an important difference between two confidence intervals as the situation where the confidence limit farthest away from 1.00, is more than twice as far away from 1.00 than the other limit is from 1.00, in absolute terms. For example, if the lower limit of the confidence interval that is farthest away from 1.00 is 1.30, then the lower limit of the other confidence interval would have to be smaller than 1.15 to constitute an important difference because 1.15 signifies point that splits the distance between 1.00 and 1.30 in two equal halves. However, when both CI limits are very close to 1.00, they will most likely yield highly similar inferences for the particular meta-analysis scenario. Thus, for this study we considered all differences between CI limits unimportant if both CI limits were between 0.90 and 1.00, or between 1.00 and 1.10.

***Exploring differences between estimates of heterogeneity***

The degree of heterogeneity between trial effect estimates is commonly assessed with the $I^2$ statistic, which measures the proportion of variance of the pooled effect due to variation between trials rather than sampling error.[22] In short, $I^2$can also be said to estimate the percentage of heterogeneity in a meta-analysis data set. $I^2$ is calculated as follows

$$I^2 = \max\left(0, \frac{Q-k+1}{Q}\right) \quad (8)$$

However, $I^2$ is just one candidate measure for the percentage of heterogeneity in a meta-analysis, and because, mathematically speaking, it is not a function of the between-trial variance, it cannot be used to compare the percentage of heterogeneity described by use of different between-trial variance estimators. An alternative way of measuring the proportion of variance of the pooled effect due to variation between trials rather than sampling error, is to utilize the assumptions that in the fixed-effect model the variation in a meta-analysis is assumed to be sampling error only, whereas under the random-effects model the variation in a meta-analysis is assumed to be a combination of sampling error and variation between trials.[22;23] The proportion of variation due to sampling error can be estimated by dividing the total variance in the fixed-effect model by the pooled variance in the random-effects model.[22;23] By subtracting this proportion from 1.0 (100%), one will have an alternative estimate of the proportion of variation between trials rather than sampling error.

Let $\hat{w}_i = \hat{\sigma}_i^2$ be the trial weights in the fixed-effect model, and $\hat{w}_i^* = 1/(\hat{\sigma}_i^2 + \hat{\tau}_E^2)$ be the trial weights in the random-effects model, where $\hat{\tau}_E^2$ is any estimate of the between-trial variance.

The pooled variance in the fixed-effect model is calculated as $v_F = 1/\left(\sum_i \hat{w}_i\right)$ and the pooled

variance in the random-effects model is calculated as $v_R = 1/\left(\sum_i \hat{w}_i^*\right)$. With respect to any between-trial variance estimator we can now calculate the percentage of heterogeneity in the meta-analysis that is not due to sampling error [23]

$$D^2 = 1 - \frac{v_F}{v_R} \tag{9}$$

That is, we calculate $D_{DL}^2$, $D_{HM}^2$, $D_{REML}^2$, $D_{HE}^2$, and $D_{SJ}^2$, when using the DL, HM, REML, HE, or SJ estimator in a random-effects model meta-analysis, respectively.

## Analysis

### Data extraction

We scanned the Cochrane Database of Systematic Reviews in The Cochrane Library, Issue 1, 2009, for *primary outcome meta-analyses*. We characterised primary outcome meta-analyses as those reporting on a binary outcome among the first three of all meta-analysed outcomes. We only included one meta-analysis per systematic review. We only included meta-analyses that included at least three trials and pooled the results from all the included trials (i.e., meta-analyses with sub-group analyses reporting only sub-totals were excluded).

### Statistical analysis

For all eligible meta-analyses we calculated the relative risks as well as the corresponding p-value, the associated 95% CI based on the conventional 'normal distribution' random-effects approach and the associated 95% CI based on the 'weighted variance' t-distribution approach.

Under the normal distribution approach we also calculated the degree of heterogeneity $D^2$ using the DL, HM, REML, HE, and SJ estimators. For trials with zero events in one intervention group we used *constant continuity correction* by adding a constant of 0.5 to the number of events and number of non-events in each intervention group.[24] As sensitivity analysis, we used *intervention group (treatment arm) continuity correction* by adding 1 divided by the number of patients in the other group to the number of events and non-events in each group.[24] As sensitivity analysis, we also repeated all analyses using odds ratio as the effect measure. Lastly, because p-values and confidence intervals depend on the precision we performed sensitivity analysis using a surrogate for the precision – the cumulative number of patients. For this analysis, we calculated the optimal information size (OIS, i.e., the required meta-analysis sample size) required to detect a relative risk reduction of 25%, based on $\alpha$=5%, $\beta$=20%, and assuming a control group risk equal the median across trials within each meta-analysis.

All analyses were performed in *R. v.2.12.*[25]

**Agreement in statistical significance**

We performed pair wise comparisons between the DL and each of the other estimators under the normal distribution approach and separately under the t-distribution approach. We created eight (4+4) 2x2 tables for the number (and percentage) of meta-analyses where the DL estimator compared to an alternative estimator (the HM, REML, HE or the SJ estimator) yielded the same or the opposite inference with regard to statistical significance. For each of these 2x2 tables we calculated the kappa value, $\kappa$, and associated 95% confidence intervals to measure agreement in statistical significance. We performed post-hoc analyses to assess how pronounced the disagreement of statistical significance was. For these analyses we grouped p-values from meta-

analyses where disagreement was observed into the four categories: 'p>0.10', 'p≤0.10 and p>0.05', 'p≤0.05 and p>0.01' and 'p≤0.01 (as explained in the previous section) and recorded the number of meta-analyses where the p-values fell into the different categories. For each comparison, we created corresponding 4x4 tables and calculated the kappa value and associated 95% confidence intervals.

**Agreement between confidence intervals in statistically significant meta-analyses**

For normal distribution based and t-distribution based meta-analyses where both the DL estimator and an alternative between-trial variance estimator yielded statistical significance (i.e., p<0.05), we plotted the 95% CI limit from the DL estimator that was closest to 1.00 (x-axis) against the 95% CI limit from the alternative estimator that was closest to 1.00 (y-axis). That is, if the meta-analysed RR was smaller than 1.00 we plotted upper 95% CI limits, and if the pooled RR was larger than 1.00 we plotted the lower 95% CI limits.

We plotted a line with slope 1, illustrating the points of complete agreement between confidence intervals. We also plotted a line with slope 0.5 and a line with slope 2, illustrating the threshold for when one CI limit was closer to 1.00 than to the other CI limit. Lastly, we counted the number of 'important' discrepancies between CI limits according to the criteria put forward in the methods section.

***Exploring differences between estimates of heterogeneity***

To compare the degree of heterogeneity arising from each estimator we constructed four plots with the estimated $D_{DL}^2$ on the x-axis, and the estimated $D_{HM}^2$, $D_{REML}^2$, $D_{HE}^2$, and $D_{SJ}^2$ respectively on the y-axis.

# Results

A total of 920 meta-analyses were eligible for our analyses.

### Agreement in statistical significance

#### *Comparisons under the normal distribution random-effects model*

Table 1 presents the number of normal distribution random-effects meta-analyses where the DL estimator, compared to the HM, REML, HE and SJ estimator yielded the same or the opposite inference with regard to statistical significance, using the constant continuity correction for handling zero-event arms.

Using the normal distribution random-effects model approach, the DL estimator yielded a statistical significance result in 414 (44.5%) of the 920 meta-analyses. Twenty eight (6.8%) of these 414 statistically significant DL meta-analyses were not statistically significant with the HM estimator, 4 (1.0%) were not statistically significant with the REML estimator, 48 (11.6%) were not statistically significant with the HE estimator, and 47 (11.4%) were not statistically significant with the SJ estimator.

The DL estimator did not yield statistical significance in 506 of the 920 meta-analyses. Two (0.4%) of these meta-analyses became statistically significant with the HM estimator, 9 (1.8%) with the REML estimator, 15 (3.0%) with the HE estimator, and 7 (1.4%) with the SJ estimator. Agreement measured with kappa-values revealed 'very good' agreement between the DL and HM estimators ($\hat{\kappa} = 0.93$, 95% CI 0.91-0.96), the DL and REML estimators ($\hat{\kappa} = 0.97$, 95% CI 0.96-0.99), the DL and HE estimators ($\hat{\kappa} = 0.86$, 95% CI 0.83-0.89), and the DL and SJ

estimators ($\hat{\kappa}$ = 0.88, 95% CI 0.85-0.91). Sensitivity analyses using 'treatment arm' continuity correction for handling zero-events, sensitivity analysis using odds ratio as the measure of effect, and sensitivity analysis by the cumulative number of patients revealed similar proportions of agreement in statistical significance and thus, similar kappa values (see tables A.1 to A.6 in the supplementary material). In the post-hoc analysis assessing how pronounced disagreements on statistical significant were, kappa values for the created 4x4 table (table 2) were slightly smaller than the kappa values for the 2x2 tables. However, all kappa values in the post-hoc analysis were larger than 80%. Of the 920 meta-analysis, 849(93%), 877(95%), 804(87%), and 811(88%) of p-values fell into the same category of strength of statistical significance when comparing the DL random-effects model to the HM, REML, HE, and SJ random-effects models, respectively. Further, 65(7%), 38(4%), 86 (9%) and 93(10%) fell into categories next to each other, and only 6(0.7%), 5(0.6%), 30(3.2%) and 16(1.7%) were two or three categories apart when comparing the DL random-effects model to the HM, REML, HE, and SJ random-effects models, respectively.

***Comparisons under the t-distribution random-effects model***

Table 3 presents the number of t-distribution random-effects meta-analyses where the DL estimator, compared to the HM, REML, HE and SJ estimator yielded the same or the opposite inference with regard to statistical significance.

The DL estimator yielded a statistical significance result in 326 (35.4%) of the 920 meta-analyses. Using the t-distribution random-effects approach, the proportions of statistically significant DL meta-analyses that were not statistically significant using an alternative estimator were approximately half or what they were under the normal-distribution random-effects

approach. The proportions of non-statistically significant DL meta-analyses that became statistically significant with an alternative estimator were between 0.7% and 1.5% across the four alternative estimators. Agreement measured with kappa-values revealed 'very good' agreement between the DL estimator and all alternative estimators as kappa estimates spanned from 0.94 to 0.97 across comparisons with lower CI limits of 0.92 or higher.

The post-hoc analysis presented in table 4 confirmed that disagreements on statistical significance are less pronounced under the t-distribution random-effects approach. In particular, kappa value estimates for all comparisons were 0.90 or larger, and the number and proportion of disagreements seen in table 4 (normal distribution) decreased by approximately 50% under the t-distribution random-effects approach.

**Agreement between confidence intervals in statistically significant meta-analyses**

***Results for the normal distribution random-effects model***

Figure 1 presents the plots of DL 95% CI limits closest to 1.00 (x-axis) against the 95% CI limits closest to 1.00 of the HM, REML, HE and SJ estimators for the normal distribution based random-effects model meta-analyses where both the DL estimator and the HM, REML, HE or SJ estimator yielded statistical significance. Overall, the differences between the 95% CIs based on HM and REML estimators and the 95% CIs based on the DL estimator were small. We counted 4 and 5 (both 1%) HM and REML meta-analyses where the differences would be considered important according to our criteria put forward in the methods section. Overall, the HM and SJ estimators were most likely to yield wider 95% CIs than that of the DL estimator but not infrequently did the opposite occur. We counted between 11 (3%) and 12 (4%) HE and SJ meta-analyses where the differences would be considered important according to our criteria put

forward in the methods section. Sensitivity analyses using 'intervention group' continuity correction for handling zero-events and sensitivity analysis using odds ratio as the measure of effect yielded similar results in the graphical inspection of CI limits as well as the counts of important discrepancies.

***Results for the t-distribution random-effects model***

Figure A.1 in the supplementary material presents the plots of DL 95% CI limits closest to 1.00 (x-axis) against the 95% CI limits closest to 1.00 of the HM, REML, HE and SJ estimators for the t-distribution based random-effects model meta-analyses where both the DL estimator and the HM, REML, HE or SJ estimator yielded statistical significance. Barely any important discrepancies were observed under the t-distribution based random-effects model.

**Exploring differences between estimates of heterogeneity**

Figure 2 presents the estimated degrees of heterogeneity (i.e., $D^2$ estimates) under the DL random-effects model meta-analyses plotted against the estimated degrees of statistical heterogeneity under the HM, REML, HE, or SJ random-effects model meta-analyses. The estimated degree of heterogeneity was almost consistently larger with the HM estimator compared to the DL estimator (a small proportion yielded up to 2% smaller estimates). The difference between the two, however, narrowed as DL estimates become larger. When the DL estimate was truncated to 0% with DL, the HM estimator yielded estimates anywhere between 0% and 90%, but most frequently below 60%. When the DL estimator yielded mild heterogeneity estimates (i.e., 30%), the HM estimator seems to yield moderate heterogeneity estimates (i.e., 30-60%).

The estimated degree of heterogeneity varied frequently between the DL and REML estimators when the DL estimator yielded estimates smaller than 60%, and especially when either of the two estimates was truncated to 0%. Absolute differences between the two were most frequently smaller than 50%. The chance of either of the two being larger than the other appears to be equal.

The estimated degree of heterogeneity varied frequently and dramatically between the DL and HE estimators and between the DL and SJ estimators for any estimate of the DL estimator. The DL estimator most frequently yielded smaller degrees of heterogeneity. Absolute differences up to 98% were observed between heterogeneity estimates from the DL estimator and either the HE or SJ estimator. Sensitivity analyses using 'treatment arm' continuity correction for handling zero-events and sensitivity analysis using odds ratio as the measure of effect yielded similar results in the graphical inspection of degrees of heterogeneity.

## Illustrative examples

In this section we provide two illustrative examples excerpted from the 920 meta-analyses included in this study. These examples illustrate how inferences about comparative effects and heterogeneity between trials may differ across random-effects meta-analyses based on different estimators and different distributional assumptions about the pooled intervention effect (normal or t-distribution).

### Example 1

A meta-analysis comparing cyclosporine with tacrolimus for preventing mortality in liver transplant patients from the systematic review 'Cyclosporin versus tacrolimus for liver

transplanted patients' yielded discrepancies in statistical significance across the five estimators, and some disagreement in the estimated degree of heterogeneity (Figure 3).[26] This meta-analysis originally found tacrolimus to be superior compared with cyclosporin (RR=0.85, 95% CI 0.73-0.99) with no heterogeneity, $I^2=D_{DL}^2=0\%$, between trials, and this finding played a dominating part in the conclusion of the systematic review.[26] Normal distribution based random-effects meta-analyses based on the HM, REML and SJ estimators did not yield statistical significance. All t-distribution based random-effects models yielded statistical significance. The HM, REML and SJ estimators produced heterogeneity estimates larger than 0% Given the relatively low proportion of events across trials and the resulting wide trial result confidence intervals, small heterogeneity estimates are expected, but should not be interpreted as ignorable.[27] The HM estimator yielded mild to moderate heterogeneity, $D_{HM}^2=32.7\%$, whereas the REML and SJ estimators yielded mild heterogeneity estimates, $D_{REML}^2=12.1\%$ and $D_{SJ}^2=13.3\%$. Inspection of the forest plot revealed some discrepancies among the larger trials. In particular, the studies by Muehlbacher et al. and O'Grady et al. seem discrepant.[26] Considerable variation in point estimates from the smaller trials were also observed. In collection, this could suggest that some degree of heterogeneity exists – inferences which could have been picked up with the HM estimator, and perhaps with the REML and SJ estimators. In the systematic review, the heterogeneity was explored by subgroup analysis of trials using oil-based cyclosporin, trials including children, trials not reporting 12 months data, trials confined to patients with hepatitis C, and trials with different protocols for immunosuppression with azathioprine or mycophenolate mofetil. No statistical evidence for subgroup effects was found. However, many of the subgroups only contained trials with a small sample and number of events, and thus, the possibility that some heterogeneity exists should not be excluded.

**Example 2**

Another meta-analysis that yielded discrepancies in statistical significance between the five estimators was the meta-analysis of corticosteroids for preventing death caused by tuberculosis meningitis from the systematic review 'Corticosteroids for managing tuberculous meningitis' (see Figure 4).[28] This meta-analysis included 7 trials, which, except for age (children, adults or both) all included similar patients. The administered interventions in the trials were similar, but the length of follow-up differed from 2 months to 2 years. The direction of effect was in favour of corticosteroids and was consistent across trials, except for the Chotmongkol et al. trial, which had imbalance in prognostic factors favouring the control (possibly due to loss to follow-up).[28] The normal distribution based DL meta-analysis yielded a relative risk of 0.79 (95% CI 0.69-0.92) and no sign of heterogeneity across trials, $I^2=D_{DL}^2=0\%$. The systematic review concluded that overall corticosteroids were effective in treating tuberculosis meningitis.[28] Normal distribution based random-effects meta-analysis using the HE and SJ estimators and t-distribution based random-effects meta-analyses using the HM, HE, and SJ estimators were not statistically significant. The HE and SJ estimators yielded very large heterogeneity estimates, $D_{HE}^2=86.7\%$ and $D_{SJ}^2=82.1\%$, and the HM estimator yielded a moderate to large heterogeneity, $D_{HM}^2=52.2\%$. Considering the consistency in effect estimates across trials, except for the Chotmongkol et al. trial, it seems more likely than not that corticosteroids are associated with prevention of death. The absence of statistical significance from the above mentioned random-effects approaches may have inappropriately downplayed the authors confidence in the overall effectiveness, had either of these approaches been used. Inferences based on the HM estimator would likely still result in the conclusion of overall effectiveness, but trigger some exploration of

heterogeneity. In this vein, one might be inclined to perform a sensitivity analysis in which the Chotmongkol et al. trial was excluded. This sensitivity analysis would reveal statistical significance and small heterogeneity estimates with all estimators.

## Discussion

Our results provide insight about the extent and frequency with which one can expect inferences from random-effects meta-analyses to vary with the choice of estimator employed to estimate the between-trial variance under the conventional normal distribution random-effects approach as well as the 'weighted variance' t-distribution approach. We explored such variations with reference to the DL estimator as this is the approach most commonly employed. We found that DL based statistical inferences about the comparative effectiveness are highly concordant with inferences based on the REML estimator. Under the normal distribution based random-effect approach, DL based statistical inferences about the comparative effectiveness were not infrequently discordant with statistical inferences based on the HM, HE and SJ estimators (roughly 5% to 10% of the included meta-analyses yielded discordant inferences). However, these inferential discordances were rarely sufficiently pronounced to impact the conclusions about the comparative treatment effect, and thus, of the systematic review conclusion. Under the t-distribution based random-effects approach, DL based statistical inferences were even less frequently discordant with statistical inferences based on any of the alternative estimators. DL based inferences about the degree of heterogeneity seem to be frequently discordant with inferences based on all four estimators. However, the types of discordances vary across estimators. The HM estimator seems predictably larger when the DL based estimates lie between 0% and 60%. The REML estimator appears smaller or larger at random by an absolute difference

up to 30% when the DL based estimates lie between 0% and 60%. The HE and SJ based estimates are, on average, larger estimates than the DL based estimates. However, differences between each of the latter two and the DL based estimates are subject to great variation and are often large.

Our two examples illustrate that the choice of between-trial variance estimator is not straight forward. In particular, they illustrate how one estimator may yield more plausible inferences for one meta-analysis data set, but less plausible inferences for another. In the face of this uncertainty, one could argue that borderline significance should not be interpreted as definitive evidence of effect – as would have been the case with example 1.

Our study offers several strengths. It represents the first large-scale empirical study on this topic, and our sample comprised of *primary outcome* meta-analyses (i.e., those most likely to influence the inferences of the systematic review). Our measures of inferential concordance and discordance cover the most commonly employed measures for statistical inference in meta-analyses (statistical significance, confidence intervals and the percentage of heterogeneity). In concert, these design features ensured high relevance to general meta-analysis and systematic review conduct.

Our study may also have potential weaknesses and limitations. We did not examine meta-analyses published in paper journals. As meta-analyses published in paper journals are more likely to present statistically significant findings and on average include more trials, one could speculate that our findings might have been different, had we included such meta-analyses.[29;30]

We only compared the DL estimator to four alternative between-trial variance estimators. There exist at least 10 additional proposals for between-trial variance estimators.[6-11;13-15;31] Thus, had we included these estimators our results might have been different – at least under the normal distribution based random-effects approach. For simplicity, we decided to obtain the relative risks (and as sensitivity analysis, odds ratios) from the eligible 920 meta-analysis data sets. However, one could argue that we should have analysed each of the 920 meta-analyses with the effects measure reported in the publication. However, such analyses would complicate comparisons of confidence intervals and heterogeneity estimates. We also did not take into account whether the authors had used a fixed-effect or random-effects model meta-analysis. However, in Cochrane meta-analyses it is uncommon to see a fixed-effect meta-analysis accompanied by a large estimate of statistical heterogeneity. Thus, p-values and 95% confidence intervals in those of the 920 meta-analysis that did employ a fixed-effect model were most likely similar to that of the DL random-effects meta-analysis. Lastly, we did not examine if any of the review authors in fact took any precautions in their assessment of statistical significance from their DL random-effects model meta-analyses. However, being regular readers of The Cochrane Library reviews – as well as other meta-analyses in paper journals – we believe such precautions are likely to be rare or non-existent.

The results from our empirical assessment are largely congruent with results from previous simulation studies. Under the normal distribution based random-effects model we observed the following about statistical inferences of the intervention effect. With the exception of a few cases, the comparative inferences for statistical significance and confidence intervals are highly similar for the DL and HM estimator. Considering the results from previous simulation studies,

this apparent similarity is, perhaps, a bit more pronounced than expected.[8;11] Although the REML estimator has generally been shown to produce superior results to the DL estimator, the differences in simulation studies have not been pronounced.[6;8;10] This was confirmed in our study. The HE and SJ estimators, on average, produce larger heterogeneity estimates.[8;10] Thus, the smaller number of statistically significant results, the wider confidence intervals, and the generally larger estimates of the percentage of heterogeneity observed in our study are no surprise. Our results also confirm that these two estimators are subject to larger variability. This added variability appears to be present for all degrees of heterogeneity, and not just when the DL estimator is truncated to 0 as previously commented on by Sidik and Jonkman.[10]

Under the t-distribution based random-effects model our results were also congruent with the findings from previous simulation studies. In particular, the inferences about the intervention effect are less affected by the choice of between-trial variance estimator under the t-distribution based random effects model.[8;11;15]

It is important to gauge the implications of the observed discrepancies between the DL estimator and alternative estimators. With regards to inferences about the meta-analysed intervention effect, the largest numbers of discrepancies were observed for the HE and SJ under the normal distribution based random-effects approach. In total, we counted 30 and 16 meta-analyses where the resulting p-values were more than two categories apart in our post-hoc analysis of agreement on statistical significance, and we counted about 13 and 12 meta-analyses where confidence intervals from significant meta-analyses disagreed (according to our arbitrary threshold). In other words no more than 43 out of 920 sampled meta-analyses (i.e., approximately 5%) were evaluated as potentially problematic. Considering the possibility that other factors could play a

dominating role in shaping the conclusion of the systematic review, checking whether the choice of between-trial variance estimator is a cause of inappropriate inferences seems less important than checking for presence of various types of bias and inadequacy of precision (i.e., insufficient number of patients and events) as causes of inappropriate inferences. One could argue that the only scenario where disclosing sensitivity to the choice of between-trial variance estimator, might be the scenario where the DL random-effects model is borderline significant and the fact that statistical significance was reached was an important determinant for the overall conclusion about comparative effectiveness.

It is also important to gauge the implications of the discrepancies between the observed degrees of heterogeneity. Many studies have demonstrated through simulation that the DL estimator, on average, underestimates heterogeneity. Consequently $I^2$ (and $D_{DL}^2$) will also be underestimates, on average. In many of the 920 meta-analyses, the $D_{DL}^2$ estimate was 0% when the other estimators yielded >0% degrees of heterogeneity. Inspection of Figure 2 could suggest that $D_{HM}^2$ provides a viable alternative to estimating the percentage of heterogeneity in a meta-analysis since this measure does not suffer from the sup-optimal truncation (to zero) property and is not subject to large variability. In contrast, $D_{HE}^2$ and $D_{SJ}^2$, although less downwardly biased, on average, are subject to large degrees of variability, and thus, presumably less reliable (this is also illustrated in example 2). Further, $D_{REML}^2$ shares the sub-optimal truncation property with $D_{DL}^2$, and does not seem to produce notably different results.

One could speculate that implementation of, for example, $D_{HM}^2$ in meta-analysis software packages and eventually standard meta-analytic practice, would result in systematic review conclusions that put more emphasis on heterogeneity. Because clinical studies are largely designed and funded in accordance with the identified gaps in the current evidence and such gaps

are frequently identified through systematic reviews, it is likely that more phase IV trials and complex phase III trials would emerge.

In summary, the estimated degree (percentage) of heterogeneity is highly sensitive to the choice of between-trial variance estimator. Inferences about the overall treatment effect, however, are only infrequently influenced by the choice of between-trial variance estimator. It is our hope that future meta-analysis software packages will incorporate measures of the degree of heterogeneity ($D^2$) based on alternative between-trial variance estimators to allow for appropriate quantification of heterogeneity. Future meta-analysis software packages should also incorporate the random-effects models based on the alternative between-trial variance estimators so that sensitivity analyses on their impact on the overall treatment effect may be carried out if necessary.

## References


(1) Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions, version 5.0.2. The Cochrane Collaboration. 2009.

(2) DerSimonian L, Laird N. Meta-analysis in clinical trials. *Contr Clin Trials* 1986; 7:177-188.

(3) Sutton A, Abrams K, Jones D, Sheldon T, Song F. Methods for meta-analysis in medical research. Chichester: Wiley; 2000.

(4) Riley RD, Higgins JP, Deeks J. Interpretation of random-effects meta-analyses. *BMJ* 2011; 342.

(5) Ioannidis JPA, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analysis. *BMJ* 2007; 335:914-916.

(6) Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. *Stat Med* 2000; 20:825-840.

(7) Brockwell SE, Gordon IR. A simple method for inference on an overall effect in meta-analysis. *Stat Med* 2007; 26:4531-4543.

(8) Sánchez-Meca J, Marín-Martínez F. Confidence intervals for the overall effect size in random-effects meta-analysis. *Psych Meth* 2008; 1:31-38.

(9) Sidik K, Jonkman JN. Simple heterogeneity variance estimation for meta-analysis. *J Roy Stat Soc (C)* 2005; 54:367-384.

(10) Sidik K, Jonkman JN. A comparison of heterogeneity variance estimators in combining resuls of studies. *Stat Med* 2007; 26:101-129.

(11) Makambi KH. The effect of the heterogeneity variance estimator on some tests of efficacy. *J Biopharm Stat* 2004; 14(2):439-449.

(12) Hartung J, Makambi KH. Reducing the number of unjustified significant results in meta-analysis. *Communications in Statistics - Simulation and Computation* 2003; 32(4):1179-1190.

(13) Hedges LV. A random-effects model for effect sizes. *Psych Bul* 1983; 93:388-395.

(14) Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *J Educa Behav Stat* 2005; 30:261-293.

(15) Sidik K, Jonkman J. Robust variance estimation for random-effects meta-analysis. *Comp Stat Data An* 2006; 50:3681-3701.

(16) Sidik K, Jonkman J. A simple confidence interval for meta-analysis. *Statistics in Medicine* 2002; 21(21):3153-3159.

(17) Sidik K, Jonkman J. On constructing confidence intervals for a standardized mean difference in meta-analysis. *Communications in Statistics - Simulation and Computation* 2003; 32:1191-1203.

(18) DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. *Contemp Clin Trials* 2007; 28:105-114.

(19) Hartung J. An alternative method for meta-analysis. *Biometrical Journal* 1999; 36:901-916.
(20) Hartung J, Knapp G. On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in Medicine* 200; 20(12):1771-1782.
(21) Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with normally distributed responses. *Statistics in Medicine* 2001; 20(24):3875-3889.
(22) Higgins JPT, Thompson S. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 2002; 21:1539-1558.
(23) Wetterslev J, Thorlund K, Brok J, Gluud C. Estimating required information size by quantifying diversity in a random-effects meta-analysis. *BMC Medical Research Methodology* 2009; 9(86).
(24) Sweeting MJ, Sutton AJ, Lampert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Stat Med* 2004; 23:1351-1357.
(25) R Core development team. R: A language and environment for statistical computing. 2010. Vienna, Austria, R Foundation for Statistical Computing.

(26) Haddad E, McAlister V, Renouf E, Malthaner R, Kjaer MS, Gluud LL. Cyclosporin versus tacrolimus for liver transplanted patients. *Cochrane Database of Systematic Reviews* 2006;(Issue 4).
(27) Rücker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I(2) in assessing heterogeneity may mislead. *BMC Medical Research Methodolgy* 2008;(8):79.
(28) Prasad K, Singh MB. Corticosteroids for managing tuberculous meningitis. *Cochrane Database of Systematic Reviews* 2008;(1):Art. No.: CD002244. DOI: 10.1002/14651858.CD002244.pub3.
(29) Moher D, Tetzlaff J, Tricco AC, et al. Epidemiology and reporting characteristics of systematic reviews. *PLoS Med* 2007; 4(3):e78.
(30) Tricco AC, Tetzlaff J, Pham B, et al. Non-Cochrane vs. Cochrane reviews were twice as likely to have positive conclusion statements: cross-sectional study. *J Clin Epi* 2009; 62(4):380-386.
(31) DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials. *Contemp Clin Trials* 7 A.D.; 28:105-114.

**Table 1** Number and percentage of normal distribution based random-effects meta-analyses where the DerSimonian-Laird (DL) estimator compared to the Hartung-Makambi (HM), restricted maximum-likelihood (REML), Hedges (HE), and Sidik-Jonkman (SJ) estimators yielded the same or opposite inference with regard to statistical significance.

| | **DL random-effects meta-analyses*** | | |
|---|---|---|---|
| **Alternative estimator** | **Not significant (n=506)** | **Significant (n=414)** | **Kappa (95%CI)** |
| ***HM random-effects meta-analyses**** | | | |
| **Non-significant** | 504 (99.6%) | 28 (7.3%) | 0.93 (0.91-0.96) |
| **Significant** | 2 (0.4%) | 386 (99%) | |
| ***REML random-effects meta-analyses**** | | | |
| **Non-significant** | 497 (98.2%) | 4 (1%) | 0.97 (0.96-0.99) |
| **Significant** | 9 (1.8%) | 410 (99%) | |
| ***HE random-effects meta-analyses**** | | | |
| **Non-significant** | 491 (97.0%) | 48 (11.6%) | 0.86 (0.83-0.89) |
| **Significant** | 15 (3.0%) | 366 (89.3%) | |
| ***SJ random-effects meta-analyses**** | | | |
| **Non-significant** | 499 (98.6%) | 47 (11.4%) | 0.88 (0.85-0.91) |
| **Significant** | 7 (1.4%) | 367 (79.5%) | |

*The percentages are calculated within the 'not significant' and 'significant' meta-analysis strata. Constant continuity correction was used for handling all zero-event arms.
CI: Confidence interval

**Table 2** Number of meta-analyses where p-values from the normal distribution based DerSimonian-Laird (DL) random-effect model compared to the normal distribution based Hartung-Makambi (HM), restricted maximum-likelihood (REML), Hedges (HE), and Sidik-Jonkman (SJ) random-effects model fell within or outside the post-hoc defined categories for the strength of statistical significance.

| | **DerSimonian-Laird p-value*** | | | | **Kappa** |
|---|---|---|---|---|---|
| | *p>0.10* | *0.10≥p>0.05* | *0.05≥p>0.01* | *p<0.01* | **(95%CI)** |
| ***Hartung-Makambi p-value**** | | | | | |
| *p>0.10* | 449 | 18 | 4 | 0 | |
| *0.10≥p>0.05* | 3 | 34 | 23 | 1 | 0.87 (0.85-0.91) |
| *0.05≥p>0.01* | 1 | 1 | 97 | 16 | |
| *p<0.01* | 0 | 0 | 4 | 269 | |
| ***Restricted maximum likelihood p-value**** | | | | | |
| *p>0.10* | 443 | 8 | 0 | 0 | |
| *0.10≥p>0.05* | 6 | 40 | 4 | 0 | 0.92 (0.90-0.94) |
| *0.05≥p>0.01* | 3 | 4 | 113 | 5 | |
| *p<0.01* | 1 | 1 | 11 | 281 | |
| ***Hedges p-value**** | | | | | |
| *p>0.10* | 445 | 12 | 12 | 8 | |
| *0.10≥p>0.05* | 4 | 30 | 26 | 2 | 0.80 (0.77-0.83) |
| *0.05≥p>0.01* | 2 | 7 | 76 | 23 | |
| *p<0.01* | 2 | 4 | 14 | 253 | |
| ***Sidik-Jonkman p-value**** | | | | | |
| *p>0.10* | 448 | 15 | 7 | 5 | |
| *0.10≥p>0.05* | 4 | 32 | 33 | 2 | 0.81 (0.78-0.85) |
| *0.05≥p>0.01* | 1 | 5 | 80 | 28 | |
| *p<0.01* | 0 | 1 | 8 | 251 | |

*Relative risk was the effect measure. 'Constant' continuity correction was used for handling all zero-event arms.
CI: Confidence interval

**Table 3** Number and percentage of t-distribution random-effects meta-analyses where the DerSimonian-Laird (DL) estimator compared to the Hartung-Makambi (HM), restricted maximum-likelihood (REML), Hedges (HE), and Sidik-Jonkman (SJ) estimators yielded the same or opposite inference with regard to statistical significance.

| | **DL random-effects meta-analyses*** | | |
|---|---|---|---|
| **Alternative estimator** | **Not significant (n=594)** | **Significant (n=326)** | **Kappa (95%CI)** |
| ***HM random-effects meta-analyses**** | | | |
| **Non-significant** | 588 (99.0%) | 13 (4.0%) | 0.96 (0.93-0.98) |
| **Significant** | 6 (1.0%) | 313 (96.0%) | |
| ***REML random-effects meta-analyses**** | | | |
| **Non-significant** | 585 (98.5%) | 2 (0.6%) | 0.97 (0.96-0.99) |
| **Significant** | 9 (1.5%) | 324 (99.4%) | |
| ***HE random-effects meta-analyses**** | | | |
| **Non-significant** | 589 (99.2%) | 16 (4.9%) | 0.95 (0.93-0.97) |
| **Significant** | 5 (0.8%) | 310 (95.1%) | |
| ***SJ random-effects meta-analyses**** | | | |
| **Non-significant** | 590 (99.3%) | 20 (6.1%) | 0.94 (0.92-0.97) |
| **Significant** | 4 (0.7%) | 306 (93.9%) | |

*The percentages are calculated within the 'not significant' and 'significant' meta-analysis strata. Constant continuity correction was used for handling all zero-event arms.
CI: Confidence interval

**Table 4** Number of meta-analyses where p-values from the t-distribution based DerSimonian-Laird (DL) random-effect model compared to the t-distribution based Hartung-Makambi (HM), restricted maximum-likelihood (REML), Hedges (HE), and Sidik-Jonkman (SJ) random-effects model fell within or outside the post-hoc defined categories for the strength of statistical significance under robust confidence intervals.

| | **DerSimonian-Laird p-value*** | | | | **Kappa** |
|---|---|---|---|---|---|
| | *p>0.10* | *0.10≥p>0.05* | *0.05≥p>0.01* | *p<0.01* | **(95%CI)** |
| ***Hartung-Makambi p-value**** | | | | | |
| *p>0.10* | 493 | 14 | 1 | 0 | |
| *0.10≥p>0.05* | 4 | 77 | 12 | 0 | 0.92 (0.90-0.94) |
| *0.05≥p>0.01* | 3 | 3 | 137 | 9 | |
| *p<0.01* | 0 | 0 | 1 | 166 | |
| ***Restricted maximum likelihood p-value**** | | | | | |
| *p>0.10* | 491 | 1 | 0 | 0 | |
| *0.10≥p>0.05* | 4 | 89 | 2 | 0 | 0.96 (0.95-0.98) |
| *0.05≥p>0.01* | 5 | 4 | 145 | 2 | |
| *p<0.01* | 0 | 0 | 4 | 173 | |
| ***Hedges p-value**** | | | | | |
| *p>0.10* | 488 | 14 | 8 | 1 | |
| *0.10≥p>0.05* | 10 | 77 | 7 | 0 | 0.90 (0.87-0.93) |
| *0.05≥p>0.01* | 1 | 3 | 134 | 9 | |
| *p<0.01* | 1 | 0 | 2 | 165 | |
| ***Sidik-Jonkman p-value**** | | | | | |
| *p>0.10* | 489 | 15 | 6 | 1 | |
| *0.10≥p>0.05* | 11 | 75 | 13 | 0 | 0.91 (0.89-0.93) |
| *0.05≥p>0.01* | 0 | 4 | 132 | 9 | |
| *p<0.01* | 0 | 0 | 0 | 251 | |

*Relative risk was the effect measure. 'Constant' continuity correction was used for handling all zero-event arms.
CI: Confidence interval

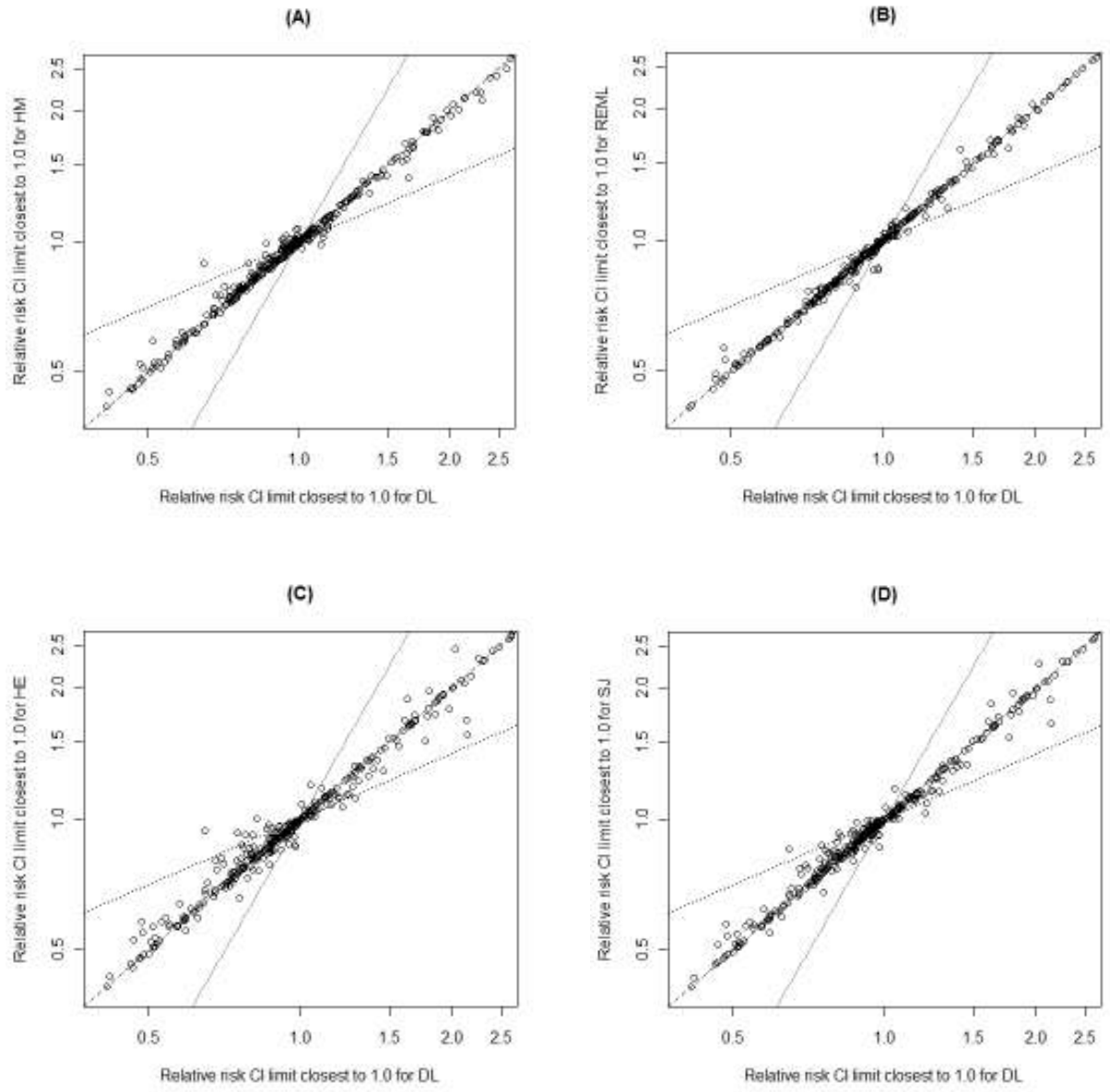



**Figure 1.** Plots of normal distribution based confidence interval (CI) limits closest to 1 (on the relative risk (RR) scale) from (A) the 410 meta-analysis where both the DerSimonian-Laird (DL) and Hartung-Makimbi (HM) random-effects models yielded statistical significance; (B) the 410 meta-analysis where both the DL and restricted maximum-likelihood (REML) random-effects models yielded statistical significance; (C) the 362 meta-analysis where both the DL and Hedges (HE) random-effects model yielded statistical significance; (D) the 329 meta-analysis where both the DL and Sidik-Jonkman (SJ) random-effects models yielded statistical significance. In all plots the CI limits from the DL random-effects model are plotted on the x-axis and the CI limits from the HM, REML, HE, and SJ random-effects model are plotted on the y-axis. Values smaller than 1.00 correspond to CI upper limits from meta-analyses where the pooled RRs are smaller than 1.00. Values larger than 1.00 correspond to CI lower limits from meta-analyses where the pooled RRs are larger than 1.00.

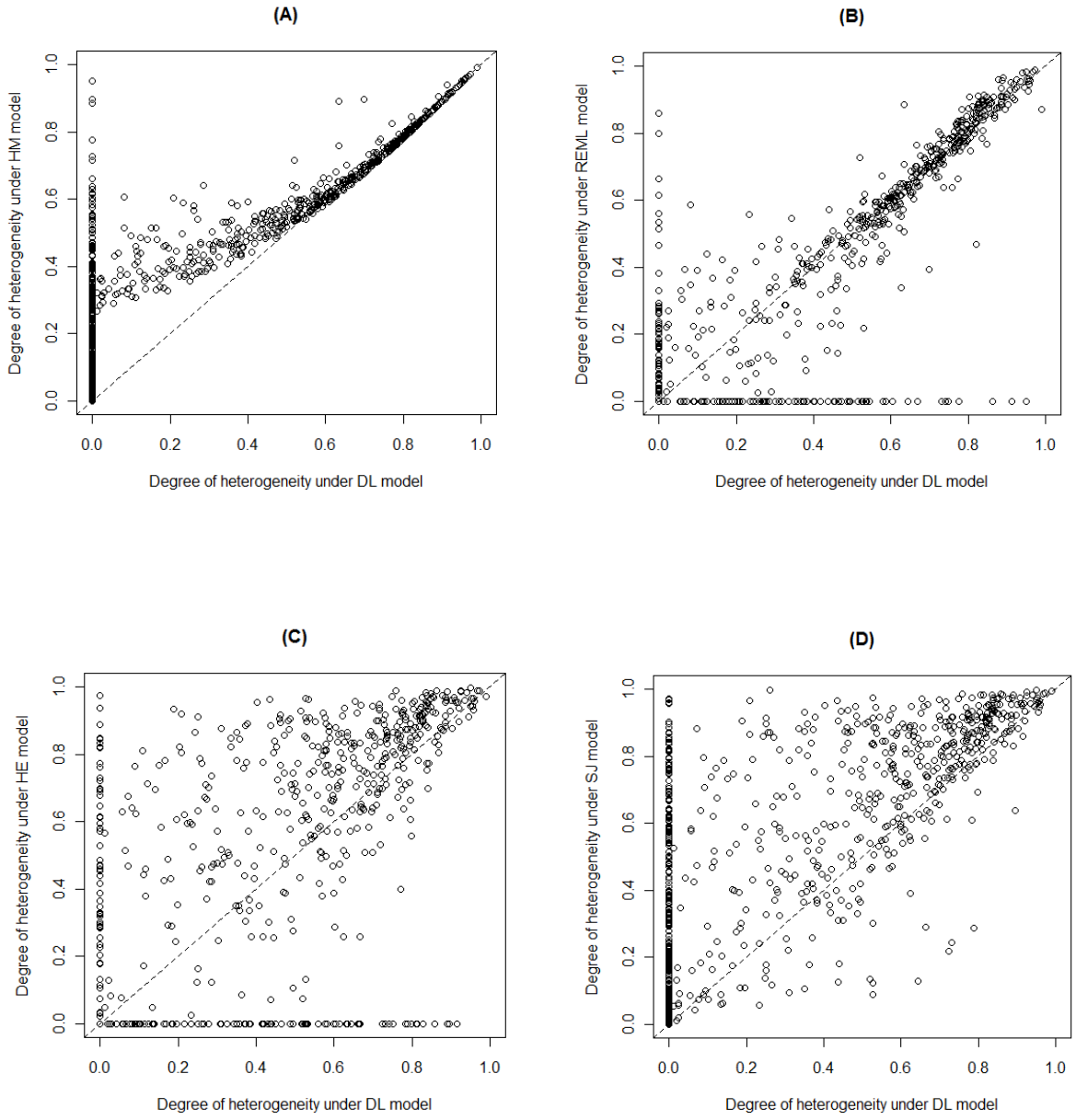
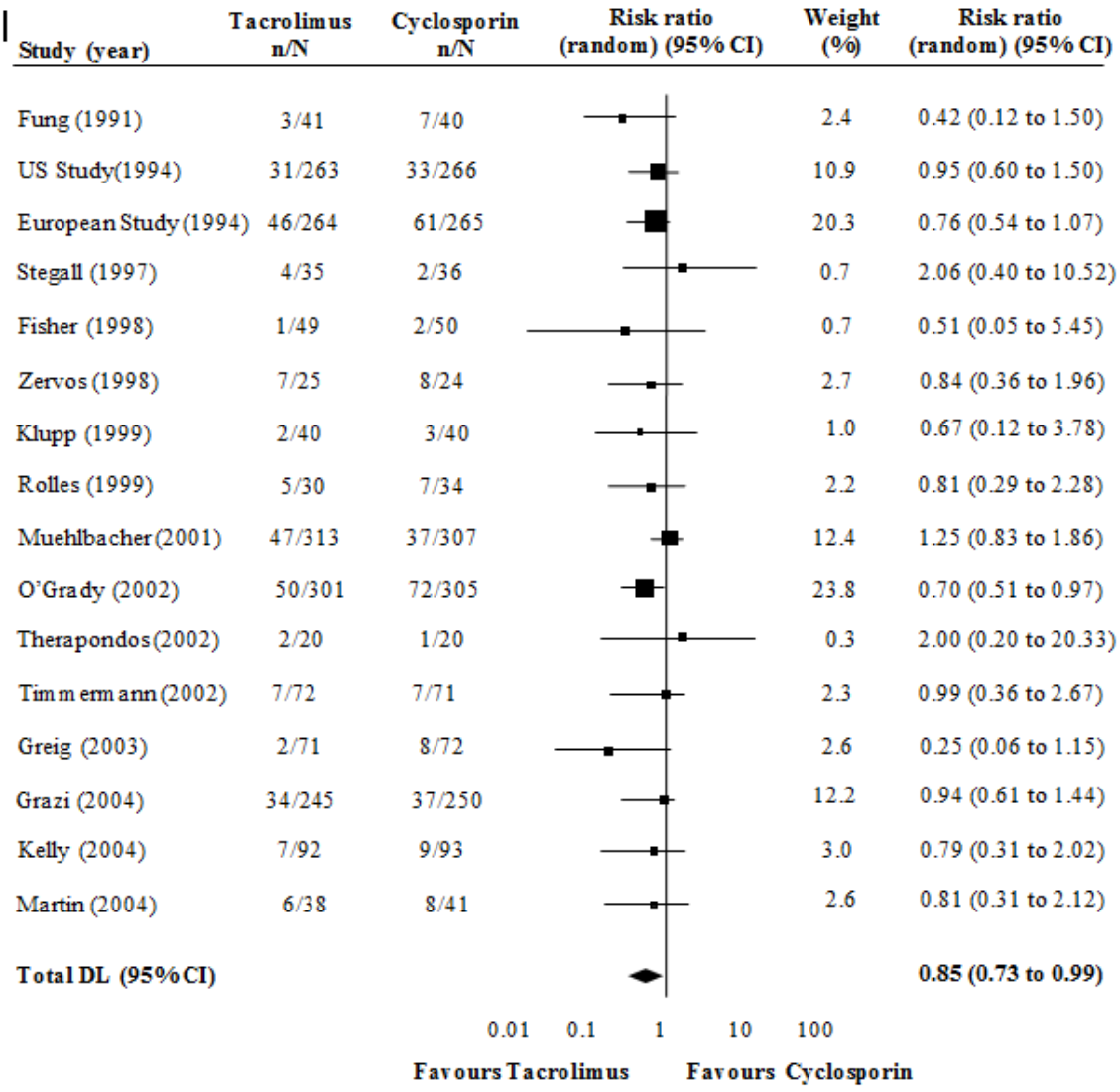
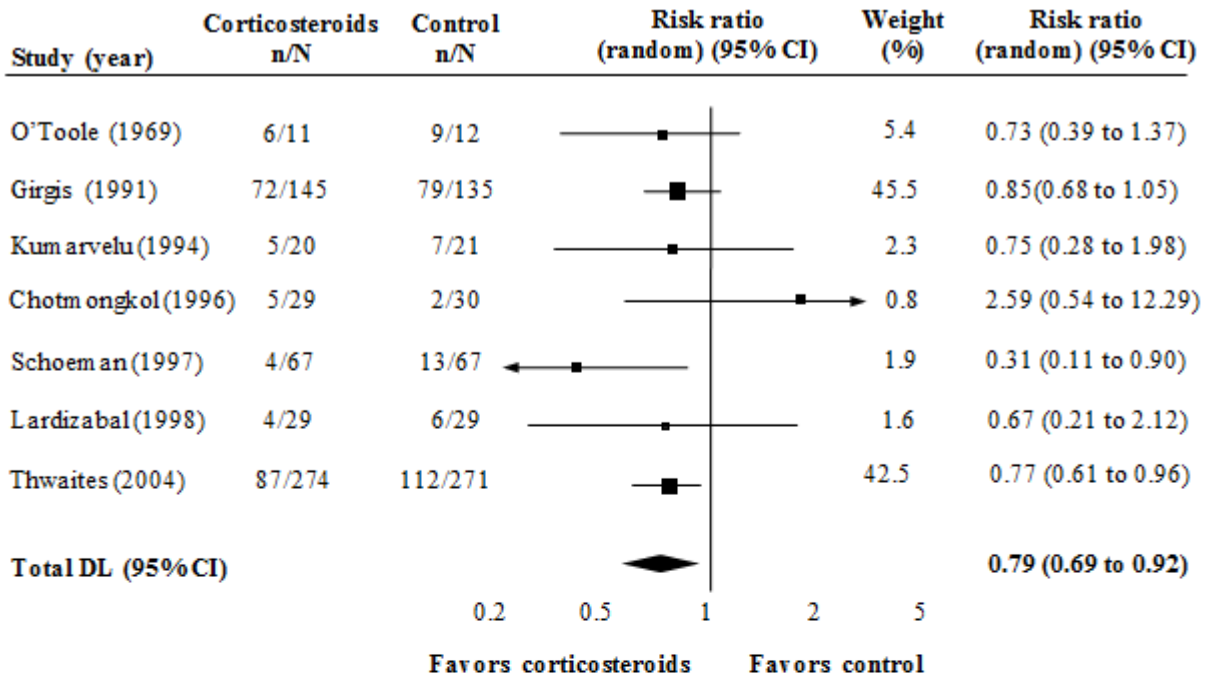



**Figure 2.** Plots of degree of heterogeneity under the DerSimonian-Laird random-effects model (x-axis) versus (A) the degree of heterogeneity under the Hartung-Makimbi (HM) random-effects models; (B) the degree of heterogeneity under the restricted maximum-likelihood (REML) random-effects models; (C) the degree of heterogeneity under the Hedges (HE) random-effects models; (D) the degree of heterogeneity under the Sidik-Jonkman (SJ) random-effects models.




| | |
|---|---|
| **DL (normal) - Test for overall effect: $P$=0.037; Heterogeneity: $D_{DL}^2$=0%** | **0.85 (0.73 to 0.98)** |
| **DL (t-dist) - Test for overall effect: $P$=0.030** | **0.85 (0.73 to 0.98)** |
| **HM (normal) – Test for overall effect: $P$=0.095; Heterogeneity: $D_{HM}^2$=32.7%** | **0.85 (0.71 to 1.03)** |
| **HM (t-dist) – Test for overall effect: $P$=0.047** | **0.85 (0.73 to 1.00)** |
| **REML (normal) - Test for overall effect: $P$=0.054; Heterogeneity: $D_{REML}^2$=12.1%** | **0.85 (0.72 to 1.00)** |
| **REML (t-dist) - Test for overall effect: $P$=0.036** | **0.85 (0.73 to 0.99)** |
| **HE (normal) - Test for overall effect: $P$=0.037; Heterogeneity: $D_{HE}^2$=0%** | **0.85 (0.73 to 0.99)** |
| **HE (t-dist) - Test for overall effect: $P$=0.030** | **0.85 (0.73 to 0.98)** |
| **SJ (normal) - Test for overall effect: $P$=0.056; Heterogeneity: $D_{SJ}^2$=13.3%** | **0.85 (0.72 to 1.00)** |
| **SJ (t-dist) - Test for overall effect: $P$=0.037** | **0.85 (0.73 to 0.99)** |

**Figure 3** Random-effects meta-analysis of tacrolimus versus cyclosporin for reducing mortality in liver transplant patients. The individual trial effect estimates and 95% confidence intervals are based on the DerSimonian-Laird (DL) random-effects model. Pooled estimates, 95% confidence intervals, P-values and percentage heterogeneity from the four alternative estimators - Hartung and Makambi (HM), restricted maximum likelihood (REML), Hedges (HE) and Sidik and Jonkman (SJ) – are presented below. For each of the estimators results under the




| | |
|---|---|
| **DL (normal) - Test for overall effect: $P$=0.002; Heterogeneity: $D_{DL}^2$=0%** | |
| **DL (t-dist) - Test for overall effect: $P$=0.02** | **0.79 (0.66 to 0.96)** |
| **HM (normal) – Test for overall effect: $P$=0.021; Heterogeneity: $D_{HM}^2$=52.2%** | **0.78 (0.63 to 0.96)** |
| **HM (normal) – Test for overall effect: $P$=0.051** | **0.78 (0.61 to 1.00)** |
| **REML (normal) - Test for overall effect: $P$=0.002; Heterogeneity: $D_{REML}^2$=0%** | **0.79 (0.69 to 0.92)** |
| **REML (t-dist) - Test for overall effect: $P$=0.02** | **0.79 (0.66 to 0.95)** |
| **HE (normal)- Test for overall effect: $P$=0.168; Heterogeneity: $D_{HE}^2$=86.7%** | **0.75 (0.50 to 1.13)** |
| **HE (t-dist)- Test for overall effect: $P$=0.135** | **0.75 (0.51 to 1.13)** |
| **SJ (normal) - Test for overall effect: $P$=0.117; Heterogeneity: $D_{SJ}^2$=82.1%** | **0.76 (0.54 to 1.07)** |
| **SJ (t-dist) - Test for overall effect: $P$=0.113** | **0.76 (0.53 to 1.09)** |

**Figure 4** Random-effects meta-analysis of corticosteroids for preventing death caused by tuberculosis meningitis. The individual trial effect estimates and 95% confidence intervals are based on the DerSimonian-Laird (DL) random-effects model. Pooled estimates, 95% confidence intervals, P-values and percentage heterogeneity from the four alternative estimators - Hartung and Makambi (HM), restricted maximum likelihood (REML), Hedges (HE) and Sidik and Jonkman (SJ) – are presented below. For each of the estimators results under the normal distribution based and t-distribution based random-effects models are presented together.

**Table A.1** Sensitivity analysis (see * below): Number and percentage of meta-analyses where the DerSimonian-Laird (DL) estimator compared to the Hartung-Makambi (HM), restricted maximum-likelihood (REML),Hedges (HE), and Sidik-Jonkman (SJ) estimators yielded the same or opposite inference with regard to statistical significance.

| | **DL random-effects meta-analyses*** | | **Kappa** |
|---|---|---|---|
| **Alternative estimator** | **Not significant (n=511)** | **Significant (n=409)** | **Estimate (95%CI)** |
| ***HM random-effects meta-analyses*** | | | |
| **Not significant** | 510 (99.9%) | 36 (8.8%) | 0.92 (0.89-0.94) |
| **Significant** | 1 (0.1%) | 373 (91.2%) | |
| ***REML random-effects meta-analyses*** | | | |
| **Not significant** | 504 (98.6%) | 8 (2%) | 0.97 (0.95-0.98) |
| **Significant** | 7 (1.4%) | 401 (98%) | |
| ***HE random-effects meta-analyses*** | | | |
| **Not significant** | 488 (95.4%) | 37 (9.0%) | 0.87 (0.84-0.90) |
| **Significant** | 23 (4.5%) | 372 (91.0%) | |
| ***SJ random-effects meta-analyses*** | | | |
| **Not significant** | 499 (97.7%) | 42 (10.3%) | 0.88 (0.85-0.91) |
| **Significant** | 12 (2.3%) | 367 (89.7%) | |

*The percentages are calculated within the 'not significant' and 'significant' meta-analysis strata. 'Treatment arm' continuity correction was used for handling all zero-event arms.
CI: Confidence interval

**Table A.2** Sensitivity analysis (see * below): Number and percentage of meta-analyses where the DerSimonian-Laird (DL) estimator compared to the Hartung-Makambi (HM), restricted maximum-likelihood (REML),Hedges (HE), and Sidik-Jonkman (SJ) estimators yielded the same or opposite inference with regard to statistical significance.

| | **DL random-effects meta-analyses*** | | |
|---|---|---|---|
| **Alternative estimator** | **Not significant (n=498)** | **Significant (n=422)** | **Kappa Estimate (95%CI)** |
| ***HM random-effects meta-analyses**** | | | |
| **Not significant** | 492 (98.8%) | 23 (5.5%) | 0.94 (0.91-0.96) |
| **Significant** | 6 (1.2%) | 399 (94.5%) | |
| ***REML random-effects meta-analyses**** | | | |
| **Not significant** | 485 (97.4%) | 6 (1.4%) | 0.96 (0.94-0.98) |
| **Significant** | 13 (2.6%) | 416 (98.6%) | |
| ***HE random-effects meta-analyses**** | | | |
| **Not significant** | 487 (97.8%) | 36 (8.5%) | 0.90 (0.87-0.93) |
| **Significant** | 11 (2.2%) | 386 (91.5%) | |
| ***SJ random-effects meta-analyses**** | | | |
| **Not significant** | 490 (98.4%) | 32 (7.6%) | 0.91 (0.89-0.94) |
| **Significant** | 8 (1.6%) | 390 (92.4%) | |

*The percentages are calculated within the 'not significant' and 'significant' meta-analysis strata. Odds ratio was the effect measure. 'Constant' continuity correction was used for handling all zero-event arms.
CI: Confidence interval

**Table A.3** Subgroup analysis by achieved levels of information size. Number and percentage of meta-analyses where the DerSimonian-Laird (DL) estimator compared to the Hartung-Makambi (HM) estimator yielded the same or opposite inference with regard to statistical significance.

| | **DL random-effects meta-analyses*** | | |
|---|---|---|---|
| **Alternative estimator** | **Not significant** | **Significant** | **Kappa (95%CI)** |
| ***Less than 50% of OIS achieved*** | | | |
| **Non-significant** | 306 (99.4%) | 18 (9.4%) | 0.91 (0.88-0.95) |
| **Significant** | 2 (0.6%) | 174 (90.6%) | |
| ***Between 50% and 100% of OIS achieved*** | | | |
| **Non-significant** | 79 (100%) | 3 (3.8%) | 0.96 (0.92-1.00) |
| **Significant** | 0 (0%) | 75 (96.2%) | |
| ***OIS surpassed*** | | | |
| **Non-significant** | 119 (100%) | 7 (4.9%) | 0.95 (0.91-0.99) |
| **Significant** | 0 (0%) | 137 (96.2%) | |

*The percentages are calculated within the 'not significant' and 'significant' meta-analysis strata. Constant continuity correction was used for handling all zero-event arms.
CI: Confidence interval

**Table A.4** Subgroup analyses by achieved levels of information size. Number and percentage of meta-analyses where the DerSimonian-Laird (DL) estimator compared to the restricted maximum likelihood (REML) estimator yielded the same or opposite inference with regard to statistical significance.

| | **DL random-effects meta-analyses*** | | |
|---|---|---|---|
| **Alternative estimator** | **Not significant** | **Significant** | **Kappa (95%CI)** |
| ***Less than 50% of OIS achieved*** | | | |
| **Non-significant** | 305 (99.0%) | 0 (0%) | 0.99 (0.97-1.00) |
| **Significant** | 3 (1.0%) | 192 (100%) | |
| ***Between 50% and 100% of OIS achieved*** | | | |
| **Non-significant** | 79 (100%) | 1 (1.3%) | 0.99 (0.96-1.00) |
| **Significant** | 0 (0%) | 77 (98.7%) | |
| ***OIS surpassed*** | | | |
| **Non-significant** | 113 (95.0%) | 3 (2.1%) | 0.93 (0.89-0.98) |
| **Significant** | 6 (5.0%) | 141 (97.9%) | |

*The percentages are calculated within the 'not significant' and 'significant' meta-analysis strata. Constant continuity correction was used for handling all zero-event arms.
CI: Confidence interval

**Table A.5** Subgroup analyses by achieved levels of information size. Number and percentage of meta-analyses where the DerSimonian-Laird (DL) estimator compared to the Hedges (HE) estimator yielded the same or opposite inference with regard to statistical significance.

| | **DL random-effects meta-analyses*** | | |
|---|---|---|---|
| **Alternative estimator** | **Not significant** | **Significant** | **Kappa (95%CI)** |
| ***Less than 50% of OIS achieved*** | | | |
| **Non-significant** | 300 (97.4%) | 20 (10.4%) | 0.88 (0.84-0.92) |
| **Significant** | 8 (2.6%) | 172 (89.6%) | |
| ***Between 50% and 100% of OIS achieved*** | | | |
| **Non-significant** | 78 (98.7%) | 10 (12.8%) | 0.86 (0.78-0.94) |
| **Significant** | 1 (1.3%) | 68 (87.2%) | |
| ***OIS surpassed*** | | | |
| **Non-significant** | 113 (95.0%) | 18 (12.5%) | 0.82 (0.75-0.89) |
| **Significant** | 6 (5.0%) | 126 (87.5%) | |

*The percentages are calculated within the 'not significant' and 'significant' meta-analysis strata. Constant continuity correction was used for handling all zero-event arms.
CI: Confidence interval

**Table A.6** Subgroup analysis by achieved levels of information size. Number and percentage of meta-analyses where the DerSimonian-Laird (DL) estimator compared to the Sidik-Jonkman (SJ) estimator yielded the same or opposite inference with regard to statistical significance.

| | **DL random-effects meta-analyses*** | | |
|---|---|---|---|
| **Alternative estimator** | **Not significant** | **Significant** | **Kappa (95%CI)** |
| ***Less than 50% of OIS achieved*** | | | |
| **Non-significant** | 308 (98.4%) | 20 (10.4%) | 0.89 (0.85-0.93) |
| **Significant** | 5 (1.6%) | 172 (89.6%) | |
| ***Between 50% and 100% of OIS achieved*** | | | |
| **Non-significant** | 79 (100%) | 9 (11.5%) | 0.89 (0.81-0.96) |
| **Significant** | 0 (0%) | 69 (88.5%) | |
| ***OIS surpassed*** | | | |
| **Non-significant** | 117 (98.3%) | 18 (12.5%) | 0.85 (0.79-0.91) |
| **Significant** | 2 (1.7%) | 126 (87.5%) | |

*The percentages are calculated within the 'not significant' and 'significant' meta-analysis strata. Constant continuity correction was used for handling all zero-event arms.
CI: Confidence interval

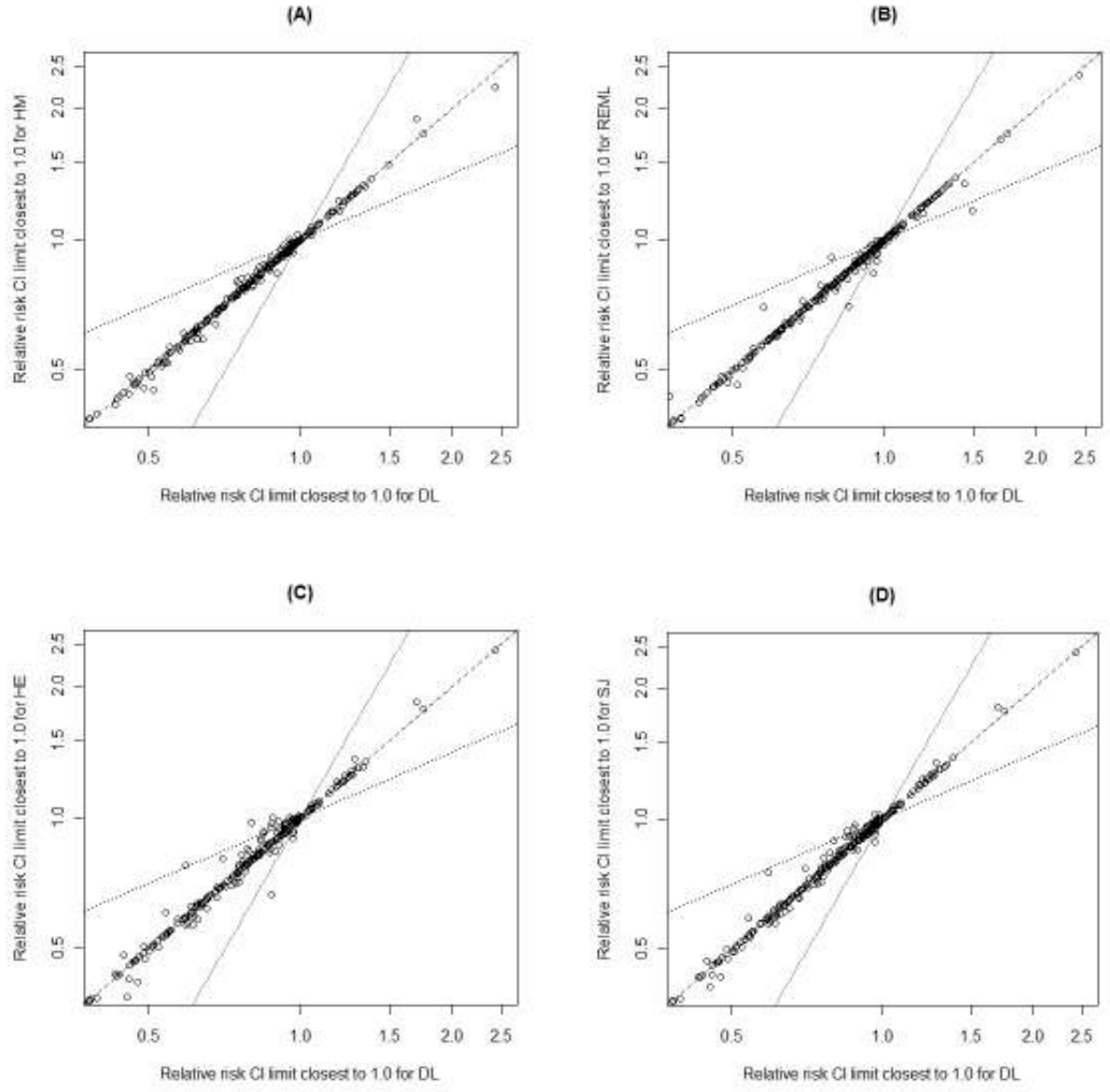



**Figure A.1.** Plots of t-distribution based confidence interval (CI) limits closest to 1 (on the relative risk (RR) scale) from (A) the 313 meta-analysis where both the DerSimonian-Laird (DL) and Hartung-Makimbi (HM) random-effects models yielded statistical significance; (B) the 324 meta-analysis where both the DL and restricted maximum-likelihood (REML) random-effects models yielded statistical significance; (C) the 310 meta-analysis where both the DL and Hedges (HE) random-effects model yielded statistical significance; (D) the 306 meta-analysis where both the DL and Sidik-Jonkman (SJ) random-effects models yielded statistical significance. In all plots the CI limits from the DL random-effects model are plotted on the x-axis and the CI limits from the HM, REML, HE, and SJ random-effects model are plotted on the y-axis. Values smaller than 1.00 correspond to CI upper limits from meta-analyses where the pooled RRs are smaller than 1.00. Values larger than 1.00 correspond to CI lower limits from meta-analyses where the pooled RRs are larger than 1.00.

# Chapter 5: Pooling continuous outcomes in meta-analysis – A tutorial and review of 12 methods for enhancing interpretability

Authors:

Kristian Thorlund

Stephen D. Walter

Bradley C. Johnston

Toshi A. Furukawa

Gordon H. Guyatt

Word count:

Summary: 200

Manuscript: 7228

## Summary

**Background:** Meta-analyses of continuous data present difficulties in interpretation when studies use different instruments to measure the same construct. Presentation of results in standard deviation units (standardized mean difference, SMD) is widely used, but is limited by vulnerability to differential variability in populations enrolled, and interpretational challenges.

**Objectives:** To identify and describe the available approaches for enhancing interpretability of meta-analyses involving continuous outcomes.

**Findings:** We identified 12 approaches in three categories:

1) Summary estimates derived from the pooled SMD: conversion to units of the most familiar instrument, or conversion to risk difference or odds ratio. These approaches remain vulnerable to differential variability in populations.

2) Summary estimates derived from the individual trial summary statistics: conversion to units of the most familiar instrument or to ratio of means. Both are appropriate complementary approaches to measures derived from converted probabilities.

3) Summary estimates derived from the individual trial summary statistics and established minimally important differences (MIDs) for all instruments: presentation in MID units or conversion to risk difference or odds ratio. Risk differences are ideal for balancing desirable and undesirable consequences of alternative interventions.

**Conclusion:** Use of these approaches may enhance the interpretability and the usefulness of systematic reviews involving continuous outcomes.

## Introduction

Meta-analyses of clinical trials typically provide enough information for decision makers (e.g. clinicians, policy makers) to evaluate the extent to which chance can explain apparent differences between interventions with respect to patient important outcomes. The interpretation of the magnitude and importance of treatment effects can, however, be challenging. When outcomes are continuous rather than dichotomous, challenges in interpretation occur as a result of two problems. First, trials often use different instruments to measure the same or similar constructs. For instance, there are at least five commonly used instruments available for measuring depression (the Beck Depression Inventory-II, Hamilton Rating Scale for Depression, Montgomery-Asberg Depression Rating Scale, Patient Health Questionnaire-9, Quick Inventory of Depressive Symptomatology.[1-5] Second, even if trials have used the same instrument, decision makers may have difficulty understanding the importance of the apparent magnitude of effect. For instance, without further information, clinicians would have difficulty grasping the importance of a 1 point difference between intervention and control in the Chronic Respiratory Questionnaire, or a 5 point difference in the SF-36.[6]

When clinical trials measure continuous outcomes involving the same or similar constructs using different instruments (e.g., depression using the Beck Depression Inventory-II and the Hamilton Rating Scale for Depression), typically review authors combine data using the *standardized mean difference* (SMD). This involves dividing the difference between the intervention and control means in each trial (i.e., the mean difference) by the estimated within-group standard deviation (SD) for that trial.[7;8] The SMD, however, has limitations. First, the SMD is measured in standard deviation units to which clinicians may not be able to easily relate.[9] Second, if the

variability or heterogeneity in the severity of patients' condition (and thus the variability in scores on the chosen outcome) varies between trials, their standard deviations will also vary. As a result, trials that enroll a heterogeneous group of patients will yield smaller SMDs than trials enrolling less heterogeneous patients, even though the unstandardized mean difference estimates - and thus the absolute estimate of the magnitude of treatment effect - may be similar across all trials.[10-12]

Many authors have proposed alternatives to the SMD that produce summary estimates that may be more easily interpreted by clinicians, some of which rely on standard deviations being similar across trials, and some of which do not.[12-16] Thus far, alternatives to the SMD have seen limited use, few studies have compared the SMD approach to the available alternatives, and a broad summary of alternative approaches, with a perspective regarding their relative merits, is unavailable.[11-14] In this article, we provide a comprehensive overview of the various methods for reporting meta-analytic summary estimates from continuous data, including their strengths and limitations. Most of the methods were originally proposed for improving interpretability of effects of individual trials, but can be readily applied to meta-analyses.

We provide a comprehensive review of 12 approaches to reporting continuous outcomes that we have grouped into three categories. The first category of approaches derive a summary estimate from the estimated meta-analysis SMD and its associated 95% confidence interval (CI). The second category of approaches first derives summary estimates for the individual trials using group means, standard deviations and sample sizes, and subsequently pools these estimates in a meta-analysis. The approaches in the third category, in addition to first deriving summary

estimates for the individual trials from group means, also depend on an established minimally important difference (MID) for the instruments employed in the relevant trials. We apply all these methods to two meta-analysis data sets: one investigating the effect of interventions for respiratory rehabilitation in chronic obstructive pulmonary disease (COPD),[17] one comparing pain in patients undergoing laparoscopic cholecystectomy receiving dexamethasone or placebo,[18] and one comparing paroxetine versus placebo for the treatment of major depression in adults.[19]

## Methods

In this section we first define some statistical notation and describe the *inverse variance method* that we use throughout our analyses to pool trial results. Because many of the approaches rely on conversion from an available mean difference estimate (or SMD) to probabilities (i.e., control and intervention group probabilities), we then establish a general methodological framework for such conversions. After presenting these general principles, we describe the identified methods ordered by category. Table 1 summarizes the strengths and limitations for each method and category.

### Statistical notation

Let $X_{Ci} \sim N(\mu_{Ci}, \sigma_{Ci}^2)$ and $X_{Ei} \sim N(\mu_{Ei}, \sigma_{Ei}^2)$ be normally distributed random variables for the responses in the control and intervention groups in trial $i$, where $\mu_{Ci}$ and $\mu_{Ei}$ are the true mean responses in the control and intervention groups in trial $i$, and $\sigma_{Ci}^2$ and $\sigma_{Ei}^2$ are the true group variances. Correspondingly let $m_{Ci}$ and $m_{Ei}$ denote the estimated mean responses, $sd_{Ci}$ and $sd_{Ei}$ denote the estimated standard deviations of $m_{Ci}$ and $m_{Ei}$, and let $n_{Ci}$ and $n_{Ei}$ denote the number of

patients in the trial groups. For trial *i*, the mean difference (MD) and its associated standard error is given by[11]

$$MD_i = m_{Ci} - m_{Ei} \quad SE(MD_i) = \sqrt{\frac{sd_{Ci}^2}{n_{Ci}} + \frac{sd_{Ei}^2}{n_{Ei}}}$$

***Pooling of results using the inverse variance method***

The pooled fixed-effect model summary estimate in a meta-analysis is typically obtained by taking a weighted average of the individual trial summary estimates. The most common type of weighted average is the *inverse variance* meta-analysis that assigns weights to individual trials according to the inverse of the variances for their summary estimates. In particular, let $\hat{\theta}_i$ denote the summary estimate of trial *i* and let $Var(\hat{\theta}_i) = \hat{\sigma}_i^2$ denote its variance (i.e., the within-trial variance). With the inverse variance method, trial *i* is then assigned the weight $w_i = 1/\sigma_i^2$ and the pooled summary estimate is derived via the formula

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i \cdot \hat{\theta}_i}{\sum_{i=1}^{k} w_i}$$

where *k* is the number of trials in the meta-analysis. The standard error of the pooled estimate, $\hat{\theta}$, is calculated as $SE(\hat{\theta}) = \sqrt{1/\sum_{i=1}^{k} w_i}$. Confidence intervals (CI) for the pooled summary estimate are typically calculated by assuming normality for $\hat{\theta}$. For example, the 95% CI is calculated as $\hat{\theta} \pm 1.96 \cdot SE\left(\hat{\theta}\right)$.

Similarly, the inverse variance method can be applied to pool results across trials under the random-effects model. The variance of each trial summary estimate, $\hat{\theta}_i$, under the random-effects model is given as the sum of the within-trial variance, $\sigma_i^2$ and the between-trial variance, $\tau_i^2$ . That is, $Var(\hat{\theta}_i) = \hat{\sigma}_i^2 + \hat{\tau}_i^2$. Using the inverse variance methods the weights under the random-model are then given by $w_i^* = \left(\hat{\sigma}_i^2 + \hat{\tau}_i^2\right)^{-1}$. The pooled estimate is calculated as

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i^* \cdot \hat{\theta}_i}{\sum_{i=1}^{k} w_i^*}$$

The standard error of the pooled estimate is given by $SE(\hat{\theta}) = \sqrt{1/\sum_{i=1}^{k} w_i^*}$ , and the 95% CIs are, again, given by $\hat{\theta} \pm 1.96 \cdot SE\left(\hat{\theta}\right)$. Further details of the conventional approach to random-effects meta-analysis are given by DerSimonian and Laird.[20]

***Conversion of individual trial continuous data to probabilities (or risks)***

Many of the methods we identified (see table 1) are based on conversion of individual trial results or the meta-analysis summary estimates into probabilities (or risks) of observing a response greater than or equal to some threshold. We first outline the general methodology for obtaining such probabilities for individual trials. In the next section we outline how these probabilities can be obtained when only the meta-analysis summary estimate is available. Note, in this paper we generally deal with situations where observing a response greater than or equal to some threshold indicates benefit, and we therefore use the word *probabilities* rather than *risks*.

Let $X_{Ci}$ and $X_{Ei}$ be normally distributed random variables for the control and intervention group mean responses in trial *i*, as defined above. For each trial, *i*, one can obtain the probabilities of observing a response greater than or equal to some threshold, *T*, in each group. For example, in trials investigating interventions for major depression, one can obtain the probability of observing a 7 point improvement from baseline or greater on the Hamilton Rating Scale for Depression. We denote the probabilities of observing a response greater than or equal to some threshold, $p_{Ci}=Pr(X_{Ci}\geq T)$ and $p_{Ei}=Pr(X_{Ei}\geq T)$. From standard normal theory these probabilities are given by

$$p_{Ci} = 1-\Phi\left(\frac{T-\mu_{Ci}}{\sigma_{Ci}}\right) \qquad p_{Ei} = 1-\Phi\left(\frac{T-\mu_{Ei}}{\sigma_{Ei}}\right) \tag{1}$$

where Φ is the standard normal cumulative distribution function.

## Category 1: Summary estimation based on the standardized mean difference

### *The standardized mean difference (SMD)*

The standardized mean difference (SMD) is the mean difference reported in standard deviation units.[7;8] It is calculated by dividing the mean difference, *MD*, by an appropriate standard deviation (e.g., the standard deviation of the control responses or the 'average' standard deviation of the control and intervention group responses). In meta-analysis, the pooled standardized mean difference is obtained by pooling individual trial standardized mean differences using the inverse variance method. That is, to calculate a pooled SMD we first calculate the individual trial SMDs and their associated standard errors and subsequently pool these using the inverse variance method.[11] Throughout this article we use standard deviations

corresponding to the weighted average of the variances of the mean responses in the two groups to standardize the individual trial mean differences. We also use the small sample adjustment for the SMD and its associated standard error. Both are incorporated in the Cochrane Collaboration's *Review Manager* software, and thus, commonly applied in practice.[11;21] The standard deviations corresponding to the weighted average of the variances of the mean responses in the two groups is calculated as[8]

$$SD(MD_i) = \sqrt{\frac{(n_{Ci}-1)sd_{Ci}^2 + (n_{Ei}-1)sd_{Ei}^2}{N_i - 2}} \qquad (2)$$

where $N_i = n_{Ci} + n_{Ei}$. The small sample adjusted SMD and its associated standard error are calculated as[8]

$$SMD_i = \frac{MD_i}{SD(MD_i)}\left(1 - \frac{3}{4N_i - 9}\right) \qquad SE(SMD_i) = \sqrt{\frac{N_i}{n_{Ci}n_{Ei}} + \frac{SMD_i^2}{2\left(N_i - 3.94\right)}}$$

All subsequent category 1 approaches rely on the pooled SMD. That is, they do make estimates from the individual trials and use these estimates to derive a pooled estimate, but then they make a direct conversion of the pooled SMD.

***i) Conversion to natural units of most familiar instrument***

One approach for improving interpretability is the conversion of a pooled SMD to a pooled mean difference in the instrument with which the target audience is most familiar. For instance, we have previously mentioned five different instruments for measuring depression. If all five had been used in different trials, one could generate an SMD and then convert to, for instance, units of the Beck Depression Scale. This conversion necessitates knowledge of the SD(MD) of the

instrument to which the author wishes to convert the SMD. For the remainder of this paper, we define the units on the most familiar instrument as *natural units*. Conversion of SMD to natural units is attained by using the formula

$$MD=SMD*SD(MD) \quad (3)$$

Confidence limits are obtained by scaling the confidence intervals of the SMD by SD(MD) (as in equation (3)). The standard deviation associated with the MD of the target instrument may be obtained in a number of ways. Using the available meta-analysis data, one could use the median, or a weighted or an unweighted average of the trial standard deviations (e.g., weight by trial sample size). One may also wish to use standard deviation estimates from external sources of evidence such as large population cohort studies. We identified two meta-analyses that have used the *conversion to natural units* approach.[22;23]

***ii-a) Conversion to probabilities and risk difference (RD) or number needed to treat (NNT)***

Risk difference (RD) and its inverse, the number needed to treat (NNT), are measures that are highly useful for trading off desirable and undesirable outcomes associated with an intervention.[24-26] To estimate the RD we must have estimates of the control and intervention group probabilities (of observing a response greater than some threshold). It is possible to obtain these probabilities from the summary SMD. The method relies on representing the meta-analysis results with conceptual projections of summary means and standard deviations for two *meta-analysis intervention groups*.[13] In the continuous meta-analysis setting in which all trials use the same instrument, we can represent the meta-analysis finding, the summary MD, with a conceptual meta-analysis control group with mean $\mu_C$, standard deviation $\sigma_C$, and group size $n_C$,

and a conceptual meta-analysis intervention group with mean $\mu_E$, standard deviation $\sigma_E$, and group size $n_C$. Thus, $MD = \mu_E - \mu_C$ and SD(MD) are appropriately derived from $\sigma_C$, $\sigma_E$, $n_C$ and $n_E$ (see equation (1)). The mean responses in both groups are assumed to follow a normal distribution.

When trials in a continuous data meta-analysis use different instruments and we choose to represent the meta-analysis finding as a summary SMD, we would construct our conceptual meta-analysis control and intervention groups so that $SMD = \mu_E - \mu_C$ and $\sigma_E = \sigma_C = 1$. Further, since conversion of the conceptual groups into probabilities is invariant to translation of the group means as long as the threshold, $T$, is translated accordingly (see equation (2)), we can set $\mu_C = 0$ for convenience. When working with standard deviation units, the threshold, $T$, can be specified indirectly in terms of some assumed conceptual control group probability, $p_C$. With $\mu_C = 0$ and $\sigma_C = 1$, we have $T=\Phi^{-1}(p_C)$, where $\Phi^{-1}$ is the inverse standard normal cumulative distribution function. We can then use the derived threshold to derive the conceptual intervention group probability. The intervention group mean response is assumed to follow a normal distribution with mean SMD and a standard deviation of 1. Thus the intervention group probability is $p_E = 1 - \Phi(T - SMD)$. Having estimated the conceptual meta-analysis control and intervention group probabilities from the pooled SMD one can derive a RD estimate using the formula[15]

$$RD = p_E - p_C \tag{4}$$

The NNT can also be estimated from the conceptual meta-analysis control and intervention group probabilities derived from the pooled SMD

$$NNT = \frac{1}{RD} = \frac{1}{p_E - p_C} \tag{5}$$

Note, both the RD and the NNT are derived directly from the pooled SMD. Confidence intervals for RD and NNT are obtained by applying the above conversion to the previously calculated confidence limits for the SMD. We identified one paper in which this method had been used to analyze meta-analysis data (i.e., obtain an NNT estimate).[27]

Although it is widely appreciated for binary data that the RD and NNT can vary considerably with the control group probability,[28] it should be noted the RD and NNT derived from the pooled SMD are very similar when the meta-analysis control probabilities s are specified as between 20% and 80%.[9]

***ii-b) Conversion to probabilities and relative risk (RR) or odds ratio (OR)***

Having specified the control group probability and calculated the intervention group probability using conversion from the pooled SMD (see ii-a), one can derive a relative risk or an odds ratio using the conventional formula[13;29]

$$RR = \frac{p_E}{p_C} \tag{6}$$

$$OR = \frac{p_E(1-p_C)}{p_C(1-p_E)} \quad (7)$$

Confidence intervals for RD and NNT are obtained by applying the above conversion to the previously calculated confidence limits for the SMD.

A simulation study demonstrated that OR estimates based on probabilities derived from the SMD may be unstable when the 'true' control group probability, $p_C$, is smaller than 20% or larger than 80%.[13], and particularly unstable when $p_C$<10% or $p_C$>90%.[13] The same study also found that this method is only preferred for converting SMD to OR (see method iii-a and iii-b below) in situations when the SMD is small and the number of trials is large.[13] Lastly, the study found that this method does not perform well when underlying distribution is skewed.[13] No simulation studies have explored the performance of RR derived from the pooled SMD. We identified one paper in which this method had been used to analyze meta-analysis data.[30]

In the discussion we elaborate on how to choose reasonable meta-analysis control group probabilities when using approaches ii-a and ii-b.

***iii-a) Conversion directly to odds ratio (OR)***

One approach that does not rely on converting the SMD to probabilities was originally proposed by Hasselblad and Hedges.[13;16] Their method exploits the fact that the logistic distribution and the standard normal distribution differ little except in the tails. The standard logistic distribution has variance $\pi^2/3$ (i.e., standard deviation $\pi/\sqrt{3}$). Therefore, a difference on the *log odds* scale can

be converted to an approximate difference on the standardized mean difference scale by dividing the difference in *log odds* by $\pi/\sqrt{3} \approx 1.81$. Similarly, an SMD can be converted to a difference in *log odds*, and thus an odds ratio

$$ln\ OR \approx -1.81 \cdot SMD$$

or

$$OR \approx e^{-1.81 \cdot SMD}$$

where *ln* is the natural logarithm. The above method assumes that the SMD is normally distributed, that the standardized group means are normally distributed with equal variances (i.e., $\sigma_E = \sigma_C = 1$), that the data follows a standard logistic distribution, and that the $p_C$ and $p_E$ from which *ln(OR)* was calculated were derived using any arbitrary threshold, *T*, (see description of approaches ii-a and ii-b). Thus, this method does not depend on nor necessitate an assumption of a single underlying conceptual meta-analysis control group probability. Under the above assumptions, the resulting OR from this method is invariant to the threshold used to dichotomize the data.[31] A comprehensive simulation study demonstrated that the direct conversion from SMD to OR typically does not perform well when the assumption of equal variances is violated or when data are skewed.[13] Further - as it is for method ii - the OR estimate becomes unstable when the 'true' control group probability is smaller than 20% or larger than 80% and in some scenarios (e.g., $p_C$ =5%) large biases can occur.[13] In other words, the resulting OR is only approximately invariant to the choice of threshold in meta-analysis scenarios in which the assumption of equal variances hold true, the data have a logistic distribution or (approximately) a normal distribution, and the threshold corresponds to a control group probability between 20% and 80%.

***iii-b) Conversion directly to OR***

Another approach for converting the SMD directly to an odds ratio recognizes that the relationship between the response probabilities for the standard logistic distribution and the cumulative probabilities for the standard normal distribution is not linear.[13;32] As noted above, the shapes of the two distributions differ most in the extreme tails, but are approximately equal elsewhere. Cox and Snell proposed an approximation based on the part of the standard logistic distribution between *p=0.20 and p=0.80*. On a plot displaying the logistic and normal distributions, Cox and Snell estimated the slope of the straight line connecting these two points and found the approximate relationship

$$ln\ OR \approx -1.65 \cdot SMD$$

or

$$OR \approx e^{-1.65 \cdot SMD}$$

This relationship holds approximately whenever the mean responses in the control and intervention approximately follow a normal distribution, and after standardization (to SD units), have equal variances. Similar to method iii-a, the OR estimate becomes unstable when the 'true' control group probability is smaller than 20% or larger than 80%.[13] Cox and Snell chose 20% and 80% because the response probabilities for the logistic distribution and the cumulative probabilities for the normal distribution are approximately linear on the logit scale between these points.[32] Beyond these points these relationships are much less linear and the above two equations are therefore inaccurate.[13;32] As for method iii-a, the OR estimate may also not perform well when the assumption of equal variances is violated or the data is skewed.[13]

Under circumstances in which estimation with approach iii-a is statistically biased (i.e., estimator bias) and where a positive SMD indicates the experimental intervention is superior, the approach is generally upwardly biased if $p_C \leq 20\%$ or $p_C \geq 80\%$, but downwardly biased if $p_C$ lies between these values.[13] Under circumstances in which estimation with approach iii-b is statistically biased and a positive SMD indicates superiority of the experimental intervention, approach iii-b is always upwardly biased, although to a much larger extent when $p_C \leq 20\%$ or $p_C \geq 80\%$.[13] Approach iii-b always yields equal or smaller standard errors compared with approach iii-a.[13] For $p_C \leq 20\%$ or $p_C \geq 95\%$, the 95% CI from approach iii-b typically does not provide the desirable coverage. For approach iii-a, the 95% CI provides desirable coverage for $p_C = 20\%$ and $p_C = 95\%$, however, when $p_C = 5\%$ the coverage is lower than the nominal level, albeit generally larger than the coverage of approach iii-a.[13]

***iv) Conversion directly to RD***

Another method for converting the SMD to a RD utilizes the relationship between these two measures and the area under the curve (AUC). For any arbitrary threshold $t=T$, let $p_C(t)$ and $p_E(t)$ be the probabilities that the standardized means in the control and intervention group, respectively, are greater than or equal to $T$. If we plot $p_C(t)$ (x-axis) against $p_E(t)$ (y-axis), we get a ROC curve. The area under this ROC curve, the AUC, is the probability that a randomly sampled patient in the intervention group has an outcome preferable to that of a randomly sampled patient in the control group. One effect measure that is strongly related to the AUC is Kraemer's 'expanded' version of the risk difference ($RD_{EXP}$): the difference between the probability that a patient in the treatment has an outcome preferable to one in the control and the

probability that a patient in the control has an outcome preferable to one in the treatment.[33;34] From this definition, the relationship between the expanded version of the risk difference and the area under the curve is given by[34;35]

$$RD_{EXP} = p_E(t) - p_C(t) = AUC - (1 - AUC) = 2 \cdot AUC - 1$$

Assuming normality and equal variances in the two groups, we have that [34;35]

$$AUC = \Phi\left(\frac{SMD}{\sqrt{2}}\right)$$

And we thus have

$$RD_{EXP} = 2 \cdot \Phi\left(\frac{SMD}{\sqrt{2}}\right) - 1$$

Again, the number needed to treat can easily be obtained based on the calculated risk difference $NNT_{EXP} = 1 / RD_{EXP}$. Confidence intervals for either RD or NNT estimates are obtained by applying the above conversion to the already obtained confidence limits for the SMD.

## Category 2: Summary estimation based on individual trial summary statistics

### i) Conversion to natural units of most familiar instrument

As noted for Category 1, approach i), when the trials included in a continuous data meta-analysis use different instruments for measuring the effect, interpretability may be enhanced by presenting results in the units of the instrument most familiar to the target audience (i.e., natural

units). Because different instruments use different scales, conversion to the units of the most popular instrument is essentially an exercise of re-scaling. The re-scaling can be conducted separately for the observed means and standard deviations in the intervention and control groups.

Assume a set of trials have used one of two instruments, A and B. Instrument A measures effects on a scale from $L_A$ to $U_A$, while instrument B uses a scale from $L_B$ to $U_B$. Assume instrument A is the more familiar, and that we want to convert mean responses and standard deviations measured by instrument B to the corresponding quantities that would have been observed if instrument A had been used. For any trial, *i*, where instrument B has been used, we have observed mean responses and standard deviations for the control group, $m_{Ci}^{B}$ and $sd_{Ci}^{B}$, and similarly $m_{Ei}^{B}$ and $sd_{Ei}^{B}$ for the intervention group. We then wish to obtain estimates, $m_{Ci}^{A}$, $sd_{Ci}^{A}$, $m_{Ei}^{A}$ and $sd_{Ei}^{A}$, of what would have been observed had instrument A been used in trial *i*. Let $R_A = U_A - L_A$ and $R_B = U_B - L_B$ be the ranges of possible values for instruments A and B, respectively. We then use the following formulas to convert (re-scale) mean estimates into the scale of instrument A

$$m_{Ci}^{A} = \left(m_{Ci}^{B} - L_B\right)\left(\frac{R_A}{R_B}\right) + L_A \quad \text{and} \quad m_{Ei}^{A} = \left(m_{Ei}^{B} - L_B\right)\left(\frac{R_A}{R_B}\right) + L_A$$

When converting mean differences that measure change (i.e., difference between pre-intervention and post-intervention results), the formula is slightly different. Let $m_{Xi-PRE}^{B}$ and $m_{Xi-POST}^{B}$ denote the pre-intervention and post-intervention results for instrument B for an intervention group *X* (where X is the group name). Then

$$m_{Xi}^{A} = \left( m_{Xi-Post}^{B} - m_{Xi-PRE}^{B} \right) \left( \frac{R_A}{R_B} \right)$$

To convert the standard deviations we simply note that standard deviations are location invariant, and so

$$sd_{Ci}^{A} = sd_{Ci}^{B} \left( \frac{R_A}{R_B} \right) \quad \text{and} \quad sd_{Ei}^{A} = sd_{Ei}^{B} \left( \frac{R_A}{R_B} \right)$$

Once all trial results have been converted to the most familiar instrument, the pooled estimate of effect (mean difference) and its associated 95% confidence interval can be obtained by simply running a conventional meta-analysis of mean differences.

**ii) Ratio of means (ROM)**

The ratio of means approach produces a relative measure of comparative effect between the control and intervention groups by dividing the mean response in the intervention group by the mean response in the control group.[14] Because the sampling distribution of the ROM is not symmetric, the individual trial ROM estimates are pooled on the log scale using the inverse variance method. For each trial, *i*, the log ROM and its associated standard error is given by

$$\ln(ROM_i) = \ln\left( \frac{m_{Ei}}{m_{Ci}} \right)$$

And

$$SE\left( \ln(ROM_i) \right) = \sqrt{\frac{1}{n_{Ei}} \cdot \left( \frac{sd_{Ei}}{m_{Ei}} \right)^2 + \frac{1}{n_{Ci}} \cdot \left( \frac{sd_{Ci}}{m_{Ci}} \right)^2}$$

A comprehensive simulation study investigated the bias, power, and confidence interval coverage associated with the ROM and found that this measure had similar statistical performance as the MD and SMD.[14] Empirical comparison of ROM to SMD and MD across 232 continuous data Cochrane meta-analyses that included at least 5 trials generally yielded concordant results for treatment effect p-values and tests and estimates of heterogeneity.[36] Linear regression with ln(ROM) as the dependent variable and SMD as the independent variable showed the following relation between the two effect measures

$$ln(ROM) = 0.392 \cdot SMD$$

with slope standard error of 0.019.[36] For SMDs of 0.2, 0.5, and 0.8, this relationship corresponds to percentage increases in the mean of 8 (95%CI 7 to 9), 22 (95%CI 19 to 24) and 37 (95%CI 33 to 41) respectively. However, this model explained only 62% of the variation (i.e., $R^2=0.62$) and it was apparent from the plot that the variation increased with the observed effect. Thus, the above confidence intervals are most likely too narrow and should be interpreted with caution. The study empirically comparing 232 continuous data Cochrane meta-analyses identified six systematic reviews that used the ROM.[36]

## Category 3: Summary estimation based on individual trials summary statistics and established minimally important differences for all instruments

### i) Mean difference in MID units

The minimal important difference is the smallest difference that, on average, patients consider important.[37-40] When a MID has been established for two or more instruments measuring the

same or similar constructs in a group of clinical trials, one can report trial mean differences in MID units, as an alternative to SD units .[12] For each trial, *i*, we divide the observed mean difference, $MD_i$ by the MID established for the instrument, *X*, used to measure the effect in trial *i*. That is, our measure of effect is $MD_i / MID_X$, where $MID_X$ denotes the MID established for instrument X. This measure has standard error

$$SE\left(\frac{MD}{MID_X}\right) = \frac{SE(MD)}{MID_X}$$

and the MID-standardized mean differences are pooled using the inverse variance method.

One limitation of reporting mean differences in MID-units is the potential errors that may arise if dissimilarities exist in measurement properties of the different instruments. Similarly, there is potential for error if the employed MIDs have been established via different methods (e.g., one MID is anchor-based whereas another is distribution-based). Lastly, mean differences reported in MID units may be susceptible to naive interpretation if emphasis is put on the pooled estimate rather than the confidence intervals. For example, a pooled MID-standardized mean difference may be smaller than 1 (i.e., on average less than a minimally important effect), but if the 95% confidence interval includes 1, one cannot rule out that the average effect may be larger than 1. Further, even if the average effect is smaller than the MID, there is still the possibility that a worthwhile proportion of patients experience an effect greater than or equal to the MID (see the following approaches for further detail).

**ii-a) Conversion to probabilities and risk difference (RD) or number needed to treat (NNT)**

When an MID has been established for all instruments used across trials, the control and intervention group responses in each trial can be converted into trial control group and intervention group probabilities by using the MID (of the respective instrument) as the threshold, *T*, in equation (2). That is, for each trial we can calculate the probability of experiencing a treatment effect greater than or equal to the MID in the control group and intervention group

$$p_{Ci} = 1 - \Phi\left(\frac{MID - m_{Ci}}{sd_{Ci}}\right) \qquad p_{Ei} = 1 - \Phi\left(\frac{MID - m_{Ei}}{sd_{Ei}}\right) \tag{8}$$

Having approximated these probabilities one can now derive the risk difference for each trial using equation (4). The standard errors for the trial risk differences are given by

$$SE(RD_i) = \sqrt{\frac{p_{Ei} \cdot (1 - p_{Ei})}{n_{Ei}} + \frac{p_{Ci} \cdot (1 - p_{Ci})}{n_{Ci}}}$$

Subsequently, one can pool the estimated trial risk differences using the inverse variance method. One can also calculate the pooled *NNT* using equation (5) (i.e., taking the inverse of the pooled RD).

**ii-b) Conversion to risks and relative risk (RR) or odds ratio (OR) using MID**

Having approximated the above trial probabilities for the two groups, one can obtain a pooled RR or pooled OR in the conventional manner. That is, first calculate the individual trial RRs or ORs (equations 6 and 7), then calculate the individual trial standard errors of the *ln* RRs or *ln*

ORs (*ln* again being the natural logarithm), the pool the logarithm transformed trial effects and exponentiate the pooled log RR or pooled log OR. The standard error of the trial *ln* RR and *ln* OR are given by

$$SE(\ln(RR_i)) = \sqrt{\frac{1}{p_{Ci} \cdot n_{Ci}} + \frac{1}{p_{Ei} \cdot n_{Ei}} - \frac{1}{n_{Ci}} - \frac{1}{n_{Ei}}}$$

$$SE(\ln(OR_i)) = \sqrt{\frac{1}{p_{Ci} \cdot n_{Ci}} + \frac{1}{p_{Ei} \cdot n_{Ei}} + \frac{1}{n_{Ci}} + \frac{1}{n_{Ei}}}$$

**ii-c) Calculating the numbers needed to treat (NNT) from the relative risk (RR) using MID**

Deriving a NNT from the pooled RD is considered sub-optimal when the control group probabilities vary appreciably across trials.[28] Under such scenarios, it is recommended that the NNT is based on the pooled RR and some control group probability which is representative of the population of interest.[28]

Having derived a pooled RR via method ii-b (using MID) and assuming some control group probability, $p_{C,}$ that is representative of a given population, one can derive an NNT for that particular population using the formula

$$NNT = \frac{1}{p_C \cdot (1 - RR)}$$

## Illustrative Examples

In all of the below example, results were pooled using the inverse variance method under the DerSimonian-Laird random-effects model.

**Example 1: Respiratory rehabilitation for COPD patients**

A Cochrane review of respiratory rehabilitation for COPD included 31 trials of which 16 employed two widely used instruments measuring disease-specific health-related quality of life: the Chronic Respiratory Disease Questionnaire (CRQ), which uses a 7-point scale (from 1 to 7), and the St. Georges Respiratory Questionnaire (SGRQ) which uses a 100-point scale (from 0 to 100).[17] Extensive evidence supports the validity and responsiveness of both these instruments, and both have established anchor-based MIDs (0.5 on the 7-point CRQ, and 4 on the 100-point SGRQ).[39;41]

We pooled data from the 16 trials that used either CRQ or SGRQ with the methods described above (Table 2). Table 3 presents the individual trial summary statistics (group mean, SD, and size). For 'conversion to natural units' (category 1-i and 2-i) we converted the results to a MD corresponding to the CRQ scale. For category 1-i, we did this by first taking the median standard deviation of the mean differences across trials that used the CRQ, and multiplied it by the SMD, as well as the calculated 95% confidence limits for the SMD. For conversion to probabilities (category 1-ii and 3-ii), we assumed a 30% 'overall' control group probability. This probability was obtained by assuming approximate normality of the control group mean responses, calculating the proportion of patients that experienced at least an MID, and using the median of proportions across all trials as an estimate for the 'overall' control group probability. We calculated the MD in MID units using the established MIDs mentioned above.

It was not appropriate to calculate the ROM because the mean differences measured change from pre-intervention to post-intervention values and some of these changes were negative.

With all methods, the summary estimates yielded large effect sizes (Table 2). All odds ratio estimates were very similar. The pooled MD in natural units (i.e., on the CRQ scale) was larger when based on the pooled SMD than when based on the individual trial summary statistics. Further, the 95%CI for the MD summary estimated based on individual trial summary statistics (0.48 to 0.94) did not include the pooled MD based on the SMD (1.02). However, both summary estimates as well as the lower 95% CI limits were larger than 0.5 – the MID for the CRQ scale. The pooled MD based on the individual trial summary statistics appeared to have higher precision. The estimate of effect in MID units was intermediate (1.75 MID units corresponds to 0.875 in natural units of the CRQ).

Summary estimates presented as RDs also showed some variation across approaches, although clinical inferences are unlikely to differ across the span of estimated RDs. The largest estimate was obtained with the direct conversion from SMD (RD=0.40, 95% CI 0.27 to 0.50), and the smallest with the conversion from SMD to probabilities (RD=0.28, 95% CI 0.19 to 0.37).

**Example 2: Prophylactic dexamethasone for nausea and vomiting after laparoscopic cholecystectomy**

A meta-analysis of prophylactic dexamethasone for nausea and vomiting after laparoscopic cholecystectomy included 17 trials of which 5 employed two widely used instruments for measuring post-operative pain: a 10-point (10 cm) Visual Analogue Scale (VAS) and a 100-point (100 mm) VAS scale.[18] Extensive evidence supports the validity and responsiveness of the VAS,[42;43] and a consensus statement suggested that 1cm on the 10cm scale constituted an MID.[44]

Although this MID is established for changes in pain, we assumed that a MID within an individual would correspond, on average, to a MID between individuals, and thus, that the 1cm MID would be applicable to post-operative pain scores.

We pooled data from the 5 trials that used two VAS scales – four of which used the 10-point scale, and one of which used the 100-point scale. The results are presented in Table 2. Table 4 presents the individual trial summary statistics (group mean, SD, and size) for the 5 trials.

For 'conversion to natural units' we converted the results to an MD corresponding to the 10-point VAS. For category 1-i, we did this by first taking the median of standard deviations for the mean differences across trials that used the 10-point VAS, and subsequently multiplying this median by the SMD. Similarly, we obtained 95% CI limits by multiplying the median SD by the 95% CI limits associated with the SMD. For conversion to probabilities (category 1-ii and 3-ii), we assumed a 20% control group risk. This was chosen based on the fact that summary estimates converted to odds ratios remain stable for control group probability between 20% and 80%. We calculated the MD in MID units using 1cm as the MID. This analysis assumes that the MID difference in post-operative pain scores is comparable to the MID for change in pain. To obtain control and intervention group probabilities derived from the MID, we assumed that for each trial the mean post-operative VAS score in the control group minus the MID (1cm) would constitute a minimally important threshold for post-operative pain. We then calculated the probabilities of patients having post-operative VAS scores larger than the MID.

Column 3 in Table 2 presents the results. The pooled mean difference in natural units (i.e., on the 10-point VAS) was larger when based on the pooled SMD (category 1) versus the individual trial summary statistics (category 2), and the 95%CI for the latter (-0.47 to -0.32) did not include the pooled summary estimate based on the SMD (-0.86). However, both summary estimates were statistically significant. The summary estimate based on individual trial summary statistics appeared to have higher precision.

Summary estimates presented as RDs were highly discrepant. The RD based on conversion to probabilities using MID was small and not statistically significant. The RD based on conversion of SMD to probabilities yielded a smaller effect than the RD based on the direct conversion from SMD, and the 95%CI of the RD based on the conversion of SMD to probabilities did not include the RD summary estimate based on the direct conversion from SMD.

The summary estimates of ORs were also discrepant across categories. The OR estimate from category 3 yielded a moderate effect which was not statistically significant, whereas the OR estimates from category 1 all yielded large statistically significant effects. The ROM summary estimate yielded a relatively small statistically significant effect. The MD presented in MID units yielded a relatively small effect which was statistically significant, the entire 95% CI being less than the minimally important difference of 1.00.

## Discussion

### Summary of findings

Meta-analyses of continuous data present difficulties in interpretation when studies use different instruments to measure the same or similar construct. Given the interpretational challenges, we have categorized and described methods for enhancing interpretability of summary estimates for continuous meta-analysis data. These methods fall into three categories based on the data and statistics from which they are derived: the pooled SMD (category 1), the individual trial summary statistics (category 2), and the individual trial summary statistics and knowledge of the MID for each instrument (category 3) (Table 1). In our examples, all approaches for obtaining odds ratios yielded similar results (Table 2). Estimates of differences (i.e., the RD, MD in natural units, and MD in MID units) were relatively similar in one examples (example 1), but discrepant in one (example 2) (Table 2).

In example 1, the observed magnitude of effect was consistently large across all three categories. The relatively large number of trials and patients, and standard deviations that were reasonably similar across trials (see Table 3) likely contribute to this consistency. Further, the instruments used in these examples (CRQ and SGRQ) also have considerable evidence of validity, are commonly used in their respective fields and have established MIDs.[39-41]

The VAS instruments employed in each of the studies in example 2 also have established measurement properties and are commonly used to measure pain.[42;43;45;46] In our dexamethasone for pain example (example 2), all summary estimates based on the pooled SMD (category 1) appeared large; whereas category 2 and 3 approaches yielded summary estimates suggesting small or moderate effects. This discrepancy most likely results from the enrolment of homogeneous populations with respect to pain, as the SDs are much smaller in relation to their

accompanying MDs than, for instance, example 1 (Tables 3 and 4). All estimates based on the pooled SMD were accompanied by a substantial degree of uncertainty (i.e., wide confidence intervals).

**Recommendations**

No single approach or category of approaches will be optimal for all continuous data meta-analysis scenarios (Table 1). A few clinical and statistical considerations can, however, facilitate the preferred approach in a given scenario. Figure 1 provides an algorithm for choosing an optimal approach to enhance interpretability. We prefer and recommend conversion to probabilities and risk difference because such measures are useful for trading off desirable and undesirable consequences.[26] Both risk differences and measures of relative effect are very familiar to clinicians and clinical researchers.

If an MID has been established for all instruments we recommend using category 3 approach ii) since this conversion is anchored in non-arbitrary thresholds (the respective instrument MIDs) and is not vulnerable to heterogeneity across SDs. We also recommend use of at least one complementary method of reporting. If some of the trials measure the effect with an instrument that is very familiar to clinicians and from which, as a result, clinicians can infer the importance of the effect, we recommend reporting results in natural units using category 2 approach i). In our examples, we have chosen the CRQ and the 10 cm VAS as the familiar instrument for natural units. We prefer the category 2 approach i) over category 1 approach i) to generating an effect in natural units because the former is not vulnerable to the heterogeneity of the SD of included studies.

If no very familiar instrument exists, but if an MID has been established for all instruments, a MD reported in MID units may be as informative as a MD reported in natural units for a very familiar instrument because clinically meaningful inferences are tied to the MID. If a familiar instrument does not exist and MIDs have not been established for all instruments, we recommend ratio of means as the complementary method of reporting since it is not vulnerable to heterogeneity of SDs.

Every approach has at least one limitation; thus our recommendations in Figure 1 should be used in concert with Table 1. For instance, we suggest that if MIDs have not been established for all instruments, category 1 approach ii will yield conversion to probabilities. This result may, however, be misleading if SDs across trials are considerably different or if, as in example 2, patients enrolled are very homogeneous, resulting in small SDs and apparently large effects. Additionally, if the preferred measure of effect is the odds ratio, direct conversion from SMD to odds ratio (category 1 approach iii) may be statistically superior to category 1, approach i) depending on the size of the effect, the sample size, the extent of variability in trial SDs, and whether outcome data are symmetrical or skewed.[13] We refer to the simulation study by Anzures et al. for further details regarding this issue.[13]

For the complementary method of reporting, Figure 1 suggests conversion to natural units (category 2 approach ii) is preferable over other methods. However, this method may not be valid if the involved instruments do not share similar measurement properties (and thus, the linear transformation is questionable) or if the validity of estimate of the SD for the most familiar

instrument is questionable. In this case, one may prefer MD in MID units or a ratio of means as the complementary method of reporting.

**A note on the use of category 3 approaches in the absence of established MIDs**

In some situations, it may be possible to use category 3 approach ii even though some of the instruments do not have an established anchor-based MID. If instruments without an established MID can be converted to an instrument with an established MID (using category 2 approach i), one could use the MID for the latter instrument on the converted summary statistics. Another approach relies on 'imputing' MIDs based on the relationship (ratio) between the MID and SD for instruments with established MIDs and using this ratio to infer the MID for instruments without an established MID.[47] For example, authors have noted that the MID is often equal to half a standard deviation.[48;49] For instruments in which this relationship has been well established, or for other reasons seems plausible, one could simply use half a standard deviation as the MID for the instruments in which the MID has not been established and carry on analyzing the data with the approached presented under category 3.[47] One could also use more sophisticated approaches for estimating the MID for instruments in which an anchor-based MID is not available (e.g. examine the relationship between the SD and MID for the instruments in which it is available and apply that ratio to the instruments for which it is not). [47]

**A note on determining control group probabilities in the absence of MIDs**

When it is not possible to derive a pooled RD, NNT, RR, or OR using MIDs, one can employ conversion to probabilities from SMD (category 1 approach ii-a, and ii-b). This approach requires specification of some control group probability that should be as non-arbitrary as

possible. In the absence of established MIDs for all instruments it may still be possible to derive meaningful plausible control group probabilities. First, if some but not all instruments have an established MID, control group probabilities can be established for these instruments using equation (8) and subsequently used to inform the value of the control group probability. Second, investigators may have reported a related dichotomous measure that one can use to specify the control group probability. For instance, in the systematic review of corticosteroids for pain, the investigators report the number of patients reporting the need for rescue analgesia.[18] Therefore, it may be advisable to establish some plausible minimum and maximum control group probability base on whatever information is available on the MIDs, and run sensitivity analysis of the chosen effect measure. Given that the derived summary estimate will typically have little sensitivity within a wide span of control group probabilities, this sensitivity analysis may yield reassuring results (i.e. the span of plausible estimates will be relatively narrow).

## Concluding remarks

We have identified and described available approaches for presenting pooled estimates of continuous data when trials measure effects using different instruments. We have summarized their relative strengths and limitations, illustrated their performance in two examples and provided recommendations for choosing optimal approaches under common clinical and statistical circumstances. In the light of our findings, we believe that the methods we have outlined and recommended are likely to enhance the interpretability, and thus, the usefulness of systematic reviews of continuous outcomes.

# References


(1) Beck AT, Steer RA, Brown GK. BDI-II. Beck Depression Inventory, 2nd ed San Antonio: Psychological Corporation, 1996; Hamilton M. Development of a rating scale for primary depressive illness. *Br J Soc Clin Psychol* 1967; 6:278-296.

(2) Kroenke K, Spitzer RL. The PHQ-9: A new depression and diagnostic severity measure. *Psychiatric Annals* 2002; 32:509-521.

(3) Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. *Br J Psychiatry* 1979; 134:382-389.

(4) Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein BN et al. The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biol Psychiatry* 2011; 54(5):573-583.

(5) Hamilton M. Development of a rating scale for primary depressive illness. *Br J Soc Clin Psychol* 1967; 6:278-296.

(6) Guyatt GH, Osoba, Wu AW, Wyrwich KW, Norman GR, Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. *Mayo Clinic Procedings* 2002; 77(4):371-383.

(7) Cohen J. Statistical power analysis for the behavioural sciences. New Jersey: Earlbaum; 198.

(8) Hedges LV, Olkin I. Statistical Methods for Meta-analysis. Orlando: Academic Press; 1985.

(9) Furukawa TA. From effect size into number needed to treat. *The Lancet* 1999; 353:1680.

(10) Greenland S, Schlesselman JJ, Criqui MG. The fallacy of employing standardized regression coefficients and correlations as measures of effect. *American Journal of Epidemiology* 1986; 123(2):203-208.

(11) Higgins JP, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions. 2009. The Cochrane Collaboration.

(12) Johnston BC, Thorlund K, Schunemann H, Xie F, Murad MH, Montori VM et al. Improving the interpretation of quality of life evidence in meta-analysis: the application of minimally important difference units. *BMC Health and Quality of Life Outcomes* 2010; 8(116):1-5.

(13) Anzures-Cabrera J, Sarpatwari A, Higgins JP. Expressing meta-analysis of continuous outcomes in terms of risks. *Statistics in Medicine* 2011.

(14) Friedrich JO, Adhikari NKJ, Beyene J. The ratio of means method as and alternative to mean differences for analyzing continuous outcome variables in meta-analysis: a simulation study. *BMC Medical Research Methodology* 2008; 8(32):1-15.

(15) Furukawa TA. From effect size into number needed to treat. *The Lancet* 1998; 353:1680.

(16) Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Quantitative Methods in Psychology* 1995; 117(1):167-178.

(17) Lacasse Y, Goldstein R, Lasserson TJ, Martin S. Pulmonary rehabilitation for chronic obstructive pulmunary disease. *Cochrane Database of Systematic Reviews* 2006; 18.

(18) Karanicolas PJ, Smith SE, Kanbur B, Davies E, Guyatt GH. The Impact of Prophylactic Dexamethasone on Vomiting and Nausea in Laparascopic Cholesystectomy Patients - A Systematic Review and Meta-analysis. *Annals of Surgery* 2008; 248:751-762.

(19) Barbui C, Furukawa TA, Cipriani A. Effectiveness of paroxetine in the acute phase treatment of adults with major depression: a systematic re-examination of published and unpublished randomised data. *CMAJ* 2008; 178:581-589.

(20) DerSimonian L, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; 7:177-188.

(21) Review Manager (RevMan) [Computer program]. Version 5.0. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2008.

(22) Lacasse Y, Wong E, Guyatt GH, King D, Cook DJ, Goldstein RS. Meta-analysis of respiratory rehabilitation in chronixc obstructive pulmonary disease. *Lancet* 1996; 348:1115-1119.

(23) Smith K, Cook D, Guyatt GH, Madhavan J, Oxman A. Respiratory muscle training in chronic airflow limitation: a meta-analysis. *Am Rev Respir Dis* 1992; 145:553-539.

(24) Aki EA, Oxman AD, Herrin J, Vist GE, Terrenato I, Sperati F et al. Using alternative statistical formats for presenting risks and risk reductions. *Cochrane Database of Systematic Reviews* 2011;(Issue 3).

(25) Guyatt GH, Rennie D, Meade MO, Cook DJ. Users' guide to the medical literature: a manual for evidence-based clinical practice. 2nd ed. New York, NY: McGraw-Hill; 2008.

(26) Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A et al. Going from evidence to recommendations. *BMJ* 2008; 336(7652):1049-1051.

(27) Busse JW, Montori VM, Krasnik C, Patelis-Siotis I, Guyatt GH, Medically Unexplained Symptoms Study Group. Psychological interventions for premenstrual syndrom: a meta-analysis of randomized trials. *Psychother Psychosom* 2009; 78:6-15.

(28) Smeeth L, Haines A, Ebrahim S. Numbers needed to treat derived from meta-analyses - sometimes informative, usually misleading. *BMJ* 1999; 338:1548-1550.

(29) Suissa S. Binary Methods for Continuous Outcomes: A Paremetric Alternative. *Journal of Clinical Epidemiology* 1991; 44(3):241-248.

(30) Furukawa TA, Akechi T, Wagenpfeil S, Leucht S. Relative indices of treatment effect may be constant across different definitions of response in schizophrenia trials. *Schizophrenia Research* 2011; 126:212-219.

(31) Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine* 2000; 19:3127-3131.

(32) Cox DR, Snell EJ. Analysis of binary data. London: Chapman and Hall; 1989.

(33) Hsu LM. Biases of success rate differences shown in binomial effect size displays. *Psychological Bulletin* 2004; 9:183-187.

(34) Kraemer HC, Kupfer DJ. Size of Treatment Effects and Their Importance to Clinical Research and Practice. *Biological Psyhology* 2006; 59:990-996.

(35) Furukawa TA, Leucht S. How to obtain NNT from Cohen's d: a comparison of two methods. *PLoS One* 2011; 6:e19070.

(36) Friedrich JO, Adhikari NKJ, Beyene J. Ratio of means for analyzing continuous outcomes in meta-analysis performed as well as mean difference methods. *Journal of Clinical Epidemiology* 2011; 64:556-564.

(37) Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR. Clinical Significance Consensus Meeting Group: Methods to explain the clinical significance of health status measures. *Mayo Clinique Procedings* 2002; 77:371-383.

(38) Jaeschke R, Singer J, Guyatt G. Measurement of Health Status: Ascertaining the Minimal Clinically Important Difference. *Controlled Clinical Trials* 1989; 10:407-415.

(39) Schunemann H, Puhan M, Goldstein R, Jaeschke R, Guyatt GH. Measurement properties and interpretability of the chronic respiratory disease questionnaire (CRQ). *COPD* 2005; 2:81-89.

(40) Schunemann H, Guyatt GH. Goodbye M(C)ID! Hello MID, where do you come from? *Health Services Research* 2005; 40(2):593-597.

(41) Jones PW, Quirk FH, Baveystock CM, Littlejohns P. A self-complete measure of health status for chronic airflow limitation. The St. George's Respiratory Questionnaire. *Am Rev Respir Dis* 1992; 145:1321-1327.

(42) Carlson AM. Assessment of chronic pain. I. Aspects of the reliability and validity of the visual analogue scale. *Pain* 1983; 16(1):87-101.

(43) DeLoach LJ, Higgins MS, Caplan AB, Stiff JL. The visual analog scale in the immediate postoperative period: intrasubject variability and correlation with a numeric scale. *Anesth Analg* 1998; 86(1):102-106.

(44) Dworkin RH, Turk DC, Wyrwich KH, et al. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. *Journal of Pain* 2009; 9:105-121.

(45) Price DD, McGrath PA, Rafii A, Buckingham B. The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. *Pain* 1983; 17(1):45-56.

(46) Wewers ME, Lowe NK. A critical review of visual analogue scales in the measurement of clinical phenomena. *Res Nurs Health* 1990; 13(4):227-236.

(47) Johnston BC, Thorlund K, da Costa BR, Furukawa TA, Guyatt GH. Improving the interpretation of quality of life evidence in meta-analyses: combining scores from instruments with and without and anchor-based minimal important difference. *Submitted* 2011.

(48) Farivar SS, Liu H, Hays RD. Half standard deviation estimate of the minimally important difference in HRQOL scores? *Expert Reviews of Pharmaeconomic Outcomes Research* 2004; 4(5):515-523.

(49) Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Medical Care* 2003; 41(5):582-592.

| Methods of reporting | Strengths | Limitations |
|---|---|---|
| ***Category 1: Summary estimation based on SMD*** | *Category strength:* Can always be calculated if variance data provided | *Category limitation:* Validity depends on standard deviations being similar across trials. |
| The standardized mean difference (SMD) | Can always be obtained | Not appealing to clinicians |
| i) Conversion to natural units of most familiar instrument | Easy to calculate if representative standard deviation of target instrument can be obtained. Easier to interpret difference on a well-known instrument than in standard deviation units. Preserves power and precision. | May provide misleading results if standard deviation of target instrument is not appropriately representative or poorly estimated. |
| ii) Conversion to probabilities and risk differences (RD), number needed to treat (NNT) or to odds ratios (OR) | Interpretation familiar to clinicians. Risk differences ideal for trading off desirable and undesirable consequences of the intervention. | May be misleading if the assumed meta-analysis control group probability is incorrect. |
| iii) Conversion directly to odds ratios | Interpretation of relative measures familiar to clinicians, though they may interpret odds ratios as relative risks. | May be inaccurate for low and high (i.e. <20% or >80%) control grou probabilities. Danger in interpreting OR as the same for all control group probabilities. |
| iv) Conversion directly to risk difference or number needed to treat | Does not depend on assumption of meta-analysis control group probability. Risk differences ideal for trading off desirable and undesirable consequences of the intervention. | Interpretation of RD and NNT somewhat different from the conventional interpretation due to derivation from 'expanded' version of RD. Danger in interpreting RD as the same for all control group probabilities. |
| ***Category 2: Summary estimation based on individual trial summary statistics*** | *Category strength:* Validity does not depend on standard deviations being similar across studies | *Category limitations:* Interpretation not tied to empirical evidence of what patients consider important. Conversion to measures that resonate well with clinicians (e.g., risk difference) not feasible. |
| i) Conversion to natural units of most familiar instrument | Easier to interpret difference on a well-known instrument than in standard deviation units. Preserves power and precision. | The linear transformation may not be valid due to different measurement properties of instruments |

| | | |
|---|---|---|
| ii) Ratio of means (ROM) | Straightforward interpretation for clinicians while retaining continuous variable measurement. Desirable statistical properties. | Not applicable to change data where mean responses can be negative |
| ***Category 3: Summary estimation based on individual trial summary statistics and established minimally important differences (MIDs) for all instruments*** | *Category strength:* Estimates intervention effects in relation to established MID. Validity does not depend on standard deviations being similar across studies | *Category limitation:* Not feasible when MID has not been established |
| i) Mean difference in MID units | Reports the mean difference on a highly interpretable scale. Preserved power and precision. | Loss of validity if dissimilarities exist in measurement properties of the different instruments, or if MIDs established via different methods (e.g., anchor vs distribution based). Susceptible to naive interpretation (e.g., effect less than 1 MID no one benefits) |
| ii) Conversion to risk and risk difference, number needed to treat or to odds ratio using MID | Interpretation familiar to clinicians. Risk differences ideal for trading off desirable and undesirable consequences of the intervention and is particularly useful if tied to the MID | Control group risk must be relatively constant for pooled RD to be meaningful. May produce artificially small effects if most patients (e.g., >95%) in both groups experience >MID effects. |

**Table 1.** Summary of strength and limitations associated with the three categories and the respective methods within each category. The category strengths and limitations apply to all methods covered under a category.

| Methods of reporting | Pooled estimate and 95% CI | |
|---|---|---|
| | **Interventions for COPD** | **Dexamethasone for pain** |
| *Category 1: Summary estimation based on SMD* | | |
| The standardized mean difference (SMD) | SMD=0.72 (95% CI 0.48 to 0.96) | SMD=-0.79 (95% CI -1.41 to -0.17) |
| i) Conversion to natural units of most familiar instrument | [1]MD=1.02 (95% CI 0.68 to 1.36) | [2]MD=-0.81 (95% CI -1.45 to -0.18) |
| ii-a) Conversion to probabilities and risk difference (RD) | [3]RD=0.28 (95% CI 0.19 to 0.37) | [4]RD=-0.15 (95% CI -0.19 to -0.04) |
| ii-b) Conversion to probabilities and odds ratio (OR) | [3]OR=3.22 (95% CI 2.24 to 4.74) | [4]OR=0.21 (95% CI 0.04 to 0.76) |
| iii-a) Conversion directly to odds ratio | OR=3.74 (95% CI 2.43 to 5.68) | OR=0.23 (95% CI 0.08 to 0.74) |
| iii-b) Conversion directly to odds ratio | OR=3.36 (95% CI 2.24 to 4.87) | OR=0.27 (95% CI 0.10 to 0.76) |
| iv) Conversion directly to risk difference | RD=0.40 (95% CI 0.27 to 0.50) | RD=-0.42 (95% CI 0.68 to -0.10) |
| *Category 2: Summary estimation based on individual trial summary statistics* | | |
| i) Conversion to natural units of most familiar instrument | [5]MD=0.71 (95% CI 0.48 to 0.94) | [6]MD=-0.35 (95% CI -0.65 to -0.05) |
| iv) Ratio of means (ROM) | [7]Not applicable | ROM=0.87 (95% CI 0.78 to 0.98) |
| *Category 3: Summary estimation based on individual trial summary statistics and established MID for all instruments* | | |
| i) Mean difference in MID units | MD=1.75 (95% CI 1.37 to 2.13) | MD=-0.40 (95% CI -0.74 to -0.07) |
| ii-a) Conversion to probabilities and RD | RD=0.31 (95% CI 0.22 to 0.40) | RD=-0.03 (95% CI -0.07 to 0.01) |
| ii-b) Conversion to probabilities and OR | OR=3.36 (95% CI 2.31 to 4.86) | OR=0.64 (95% CI 0.34 to 1.17) |

[1]Based on the median SD=1.32 for the CRQ scale. [2]Based on the median SD=1.04 for the 10-point VAS scale. [3]Based on 30% control group risk (which is approximately the median of the 16 trials). [4]Based on 20% control group risk (sensitivity analysis using control groups risks between 10% and 90% yielded similar findings). [5]Measured on CRQ scale. [6]Measured on a 10-point Visual Analogue Scale (VAS). [7]Measure change from baseline which can be negative and therefore has no natural zero

**Table 2.** Summary estimates and their associated 95% confidence intervals from each of the identified methods applied to the two data sets.

| | Intervention group | | | Control group | | |
|---|---|---|---|---|---|---|
| | **Mean** | **SD** | **n** | **Mean** | **SD** | **n** |
| ***Trials that used CRG*** | | | | | | |
| Behnke (2000) | 1.90 | 0.70 | 15 | -0.07 | 1.10 | 15 |
| Cambach (1997) | 1.04 | 0.91 | 15 | 0.01 | 0.75 | 8 |
| Goldstein (1994) | 0.43 | 0.92 | 40 | -0.13 | 0.75 | 40 |
| Gosselink (2000) | 0.67 | 1.02 | 34 | -0.10 | 1.11 | 28 |
| *Griffiths (2000) | 0.97 | 1.00 | 93 | -0.15 | 0.90 | 91 |
| Guell (1995) | 0.98 | 1.01 | 29 | -0.18 | 1.05 | 27 |
| Guell (1998) | 0.45 | 0.89 | 18 | -0.30 | 0.97 | 17 |
| Hernandez (2000) | 0.86 | 1.00 | 20 | 0.14 | 1.03 | 17 |
| Simpson (1992) | 0.86 | 1.26 | 14 | 0.13 | 1.11 | 14 |
| Singh (2003) | 0.91 | 0.75 | 20 | 0.10 | 0.68 | 20 |
| Wijkstra (1994) | 0.80 | 0.83 | 28 | 0.07 | 0.82 | 15 |
| ***Trials that used SGRQ*** | | | | | | |
| Boxall (2005) | -5.8 | 11.8 | 23 | -1.4 | 13.3 | 23 |
| Chlumsky (2001) | -4.1 | 19.8 | 13 | -4.2 | 19.2 | 6 |
| Engstrom (1999) | 0.3 | 17.3 | 26 | 0.5 | 16.2 | 24 |
| Finnerty (2001) | -9.3 | 12.2 | 24 | -2.2 | 15.0 | 25 |
| *Griffiths (2000) | -7.1 | 15.5 | 93 | 1.3 | 11.7 | 91 |
| Ringbaek (2000) | -2.1 | 19.0 | 17 | -2.2 | 17.0 | 19 |

* Griffiths study used both CRQ and SGRQ. Only CRQ results were used for the analyses.

**Table 3.** Individual trial summary statistics from the meta-analysis on interventions for COPD

| | Intervention group | | | Control group | | |
|---|---|---|---|---|---|---|
| | **Mean** | **SD** | **n** | **Mean** | **SD** | **N** |
| ***Trials that used 10-point VAS*** | | | | | | |
| Elkahim (2002) | 3.06 | 1.24 | 120 | 3.70 | 1.40 | 30 |
| Feo (2005) | 1.71 | 0.44 | 49 | 1.83 | 0.29 | 52 |
| Nesek-Adam (2007) | 2.76 | 0.57 | 80 | 2.91 | 0.69 | 80 |
| Wang (1999) | 2.65 | 0.77 | 40 | 2.90 | 0.89 | 38 |
| ***Trials that used 100-point VAS*** | | | | | | |
| Ozdamar (2006) | 20.4 | 2.31 | 25 | 27.1 | 1.86 | 25 |

**Table 4.** Individual trial summary statistics from the meta-analysis on dexamethasone for reducing post-operative pain in patients undergoing laparoscopic cholecystectomy.

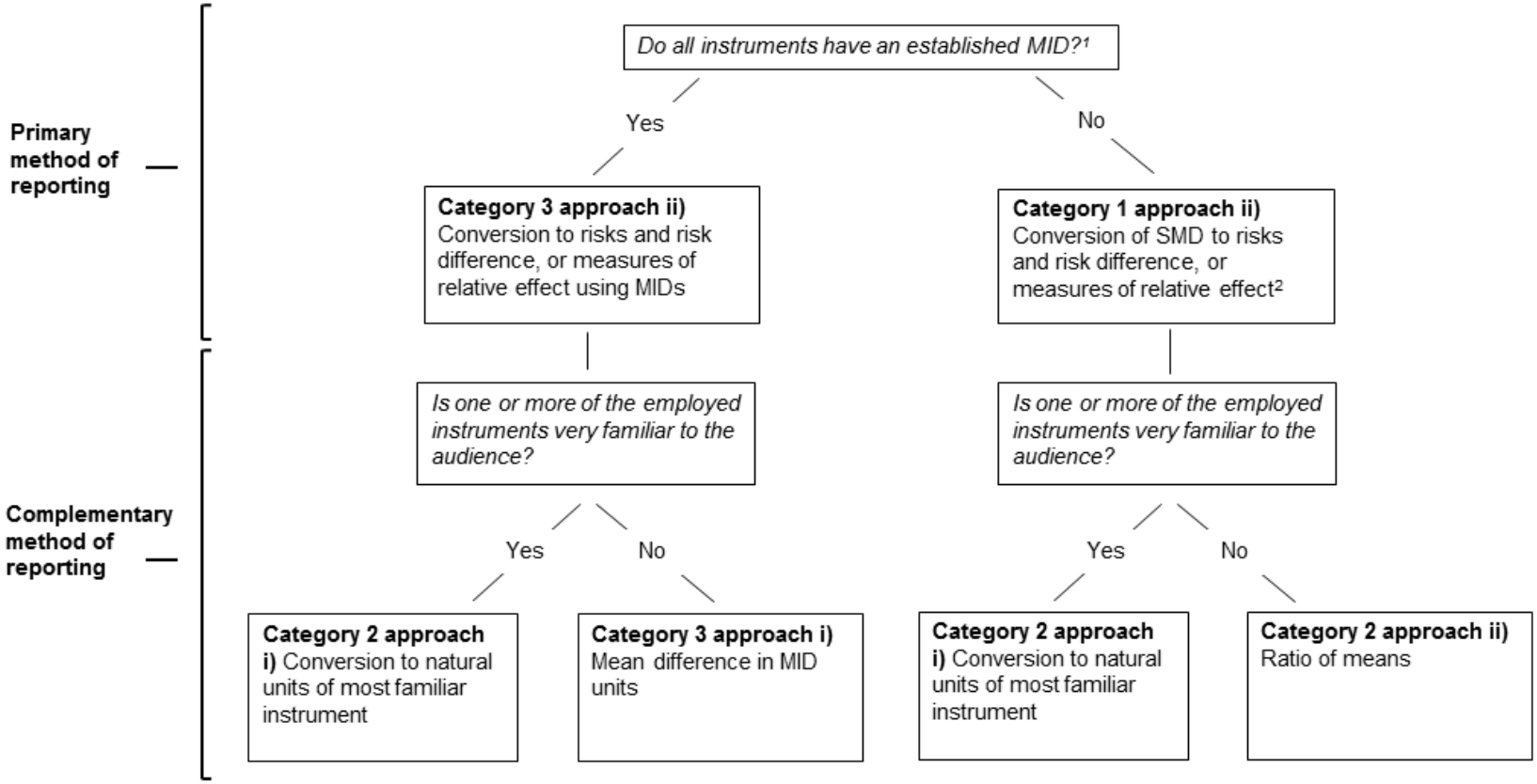



**Figure 1** Recommendations for choosing a statistical method to enhance interpretability. We recommend presenting results through risks as the primary method of reporting and supplementing with at least one complementary approach depending on what method is most likely to resonate well with the intended audience. We recommend using the above flow chart in concert with table 1.

[1] An established MID may not be required for all instruments if statistical approaches to estimating the MID for instruments in which it has not been established can be applied in a valid fashion (see text).
[2] Category 1, approach iii may be preferable to approach ii under certain circumstances depending on the size of the effect, the sample size, the extent of variability in trial standard deviations, and whether outcome data are symmetrical or skewed (see text).

# Chapter 6: Some additional analyses linking the issues explored in chapters 2, 3 and 4

## Additional analysis #1:

## Stability of alternative measures of the percentage of heterogeneity ($D^2$)

In chapter 3, it was explored how the degree of heterogeneity, $I^2$, evolved (and fluctuated) as the number of events and trials increased. In chapter 4, an alternative method for measuring the degree of heterogeneity, $D^2$, was employed because this measure allowed for comparison of the degree of heterogeneity under all five random-effects models (between-trial variance estimators) that were studied in this chapter. Given that the $I^2$ estimate demonstrated considerable variation over time in chapter 3, it seems worthwhile to explore whether any of the five alternatives considered in chapter 4 are less variable over time and become stable with fewer events and trials.

Figures 1 to 10 show the plots of the cumulative $I^2$ estimate and the cumulative $D^2$ estimate based on each of the five between-trial variance estimators considered in chapter 4. In addition to plotting the cumulative $I^2$ estimate, Figures 1 and 2 plot the cumulative $D_{DL}^2$ estimate, Figures 3 and 4 plot the cumulative $D_{HM}^2$ estimate, Figures 5 and 6 plot the cumulative $D_{REML}^2$ estimate, Figures 7 and 8 plot the cumulative $D_{HE}^2$ estimate, and Figures 9 and 10 plot the cumulative $D_{SJ}^2$ estimate.
The cumulative $D_{DL}^2$ estimate is consistently greater than or equal to the cumulative $I^2$ estimate. Fluctuations in these two measures always occur in the same direction, and it seems that $D_{DL}^2$ consistently incurs greater or similar fluctuations than $I^2$. The cumulative $D_{HM}^2$ estimate is consistently greater than the cumulative $I^2$ estimate. In roughly half of the 16 meta-analyses, $D_{HM}^2$ seems to stabilize faster or incur less fluctuations compared with $I^2$, in the other half the two measures incur approximately the same fluctuation, albeit often in different directions. In most of

the 16 meta-analyses, $D_{REML}^2$, $D_{HE}^2$, and $D_{SJ}^2$ all incur moderate to substantially larger fluctuations than $I^2$ or stabilize slower.

In summary, $D_{HM}^2$ appears to be the preferred measure of heterogeneity. This additional analysis thereby confirms the recommended use of $D_{HM}^2$ in chapter 4.

## Additional analysis #2:

## Discrepancies between random-effects model inferences in relation to the optimal information size (OIS)

In chapter 2, I demonstrated that the risk and possible magnitude of overestimation of meta-analyzed intervention effects were generally 'acceptably' low once a meta-analysis had surpassed its OIS, and often considerably before that. As mentioned in chapter 3, the reliability of heterogeneity estimates is to some degree determined by the reliability of the meta-analysed intervention effect. This is also the case for the reliability of all between-trial variance estimators considered in chapter 4 since they are all a function of the sum of squared distances between the trial intervention effect estimates and the meta-analysed intervention effect estimate. An unreliable meta-analysed intervention effect estimate could impact the between-trial variance estimate. Depending on which estimator is used, this impact may be small or large and the impact may depend on the accumulated amount of information.

In chapter 4, basic statistical inferences from five random-effects models (based on five different between-trial variance estimators) were compared among 920 Cochrane 'primary outcome' meta-analyses. Not infrequently were discrepancies observed with regard to conventional statistical significance for some random-effects models (estimators), and quite often the measured degree of heterogeneity varied between models. Given the findings in chapters 2 and 3 and the above additional analysis, it therefore seemed worthwhile to explore whether the intensity of inferential discrepancies observed in chapter 4 would vary with the meta-analysis information size. In this additional analysis the primary analyses from chapter 4 on statistical significance and degrees of heterogeneity are repeated for three subgroups of the 920 meta-analyses. Each of the three

subgroups are determined by the accumulated meta-analysis information size: 'less than 50% of the OIS', '50% to 100% of the OIS', and 'OIS surpassed'. The OIS estimates for all meta-analyses were geared to detect a relative risk reduction of 25% with a maximum 5% type I error, maximum 10% type II error (i.e., 90% power), assuming a control group risk equal to the median among all trials, and correcting for a 33% heterogeneity. Due to a comment from a peer reviewer in between a previous version and the final version of this thesis, some of these analyses are already included chapter 4.

Tables 1 to 4 show the number and percentage of disagreements on conventional statistical significance sub-grouped by the information sizes. The inferences from the DerSimonian-Laird (DL) random-effects model are compared to the inferences from the Hartung-Makambi (HM) random-effects model in Table 1, the restricted maximum likelihood (REML) random-effects model in Table 2, the Hedges (HE) random-effects model in Table 3, and the Sidik-Jonkman (SJ) random-effects model in Table 4. For the comparison of statistical significance inferences of the DL and HM random-effects models, the kappa estimate improves slightly when going from less than 50% of the OIS to between 50% and 100% of the OIS. As well, the percentage of statistically significant DL random-effects meta-analyses which are not statistically significant with the HM random-effects model is roughly twice as large when less than 50% of the OIS is accumulated. For the remaining comparisons, a slight decrease in the kappa estimate occurs for meta-analyses where the OIS is surpassed. This seems to be due to an increase in non-significant DL meta-analyses which become statistically significant when using any of the three alternative models (REML, HE, or SJ).

Figures 11 to 14 present the degree of heterogeneity under the DerSimonian-Laird random-effects model (x-axis) plotted against the degree of heterogeneity under the each of the four alternative

models (y-axis). No apparent differences were observed in the plots. Thus, on average, discrepancies about the degree of heterogeneity seem unaffected by the amount of accumulated information.

In summary, inferential discrepancies caused by the choice of between-trial variance estimator, on average, appear relatively constant with varying accumulated meta-analysis information.

# Tables

| Alternative estimator | DL random-effects meta-analyses* Not significant (n=324) | Significant (n=192) | Kappa (95%CI) |
|---|---|---|---|
| ***Less than 50% of OIS achieved*** | | | |
| **Non-significant** | 306 (99.4%) | 18 (9.4%) | 0.91 (0.88-0.95) |
| **Significant** | 2 (0.6%) | 174 (90.6%) | |
| ***Between 50% and 100% of OIS achieved*** | | | |
| **Non-significant** | 79 (100%) | 3 (3.8%) | 0.96 (0.92-1.00) |
| **Significant** | 0 (0%) | 75 (96.2%) | |
| ***OIS surpassed*** | | | |
| **Non-significant** | 119 (100%) | 7 (4.9%) | 0.95 (0.91-0.99) |
| **Significant** | 0 (0%) | 137 (96.2%) | |

*The percentages are calculated within the 'not significant' and 'significant' meta-analysis strata. Constant continuity correction was used for handling all zero-event arms.
CI: Confidence interval

**Table 1** Subgroup analysis by achieved levels of information size. Number and percentage of meta-analyses where the DerSimonian-Laird (DL) estimator compared to the Hartung-Makambi (HM) estimator yielded the same or opposite inference with regard to statistical significance.

| | DL random-effects meta-analyses* | | |
|---|---|---|---|
| **Alternative estimator** | **Not significant (n=324)** | **Significant (n=192)** | **Kappa (95%CI)** |
| ***Less than 50% of OIS achieved*** | | | |
| **Non-significant** | 305 (99.0%) | 0 (0%) | 0.99 (0.97-1.00) |
| **Significant** | 3 (1.0%) | 192 (100%) | |
| ***Between 50% and 100% of OIS achieved*** | | | |
| **Non-significant** | 79 (100%) | 1 (1.3%) | 0.99 (0.96-1.00) |
| **Significant** | 0 (0%) | 77 (98.7%) | |
| ***OIS surpassed*** | | | |
| **Non-significant** | 113 (95.0%) | 3 (2.1%) | 0.93 (0.89-0.98) |
| **Significant** | 6 (5.0%) | 141 (97.9%) | |

*The percentages are calculated within the 'not significant' and 'significant' meta-analysis strata. Constant continuity correction was used for handling all zero-event arms.
CI: Confidence interval

**Table 2** Subgroup analyses by achieved levels of information size. Number and percentage of meta-analyses where the DerSimonian-Laird (DL) estimator compared to the restricted maximum likelihood (REML) estimator yielded the same or opposite inference with regard to statistical significance.

| | DL random-effects meta-analyses* | | |
|---|---|---|---|
| **Alternative estimator** | **Not significant (n=324)** | **Significant (n=192)** | **Kappa (95%CI)** |
| ***Less than 50% of OIS achieved*** | | | |
| **Non-significant** | 300 (97.4%) | 20 (10.4%) | 0.88 (0.84-0.92) |
| **Significant** | 8 (2.6%) | 172 (89.6%) | |
| ***Between 50% and 100% of OIS achieved*** | | | |
| **Non-significant** | 78 (98.7%) | 10 (12.8%) | 0.86 (0.78-0.94) |
| **Significant** | 1 (1.3%) | 68 (87.2%) | |
| ***OIS surpassed*** | | | |
| **Non-significant** | 113 (95.0%) | 18 (12.5%) | 0.82 (0.75-0.89) |
| **Significant** | 6 (5.0%) | 126 (87.5%) | |

*The percentages are calculated within the 'not significant' and 'significant' meta-analysis strata. Constant continuity correction was used for handling all zero-event arms.
CI: Confidence interval

**Table 3** Subgroup analyses by achieved levels of information size. Number and percentage of meta-analyses where the DerSimonian-Laird (DL) estimator compared to the Hedges (HE) estimator yielded the same or opposite inference with regard to statistical significance.

| | DL random-effects meta-analyses* | | |
|---|---|---|---|
| **Alternative estimator** | **Not significant (n=324)** | **Significant (n=192)** | **Kappa (95%CI)** |
| ***Less than 50% of OIS achieved*** | | | |
| **Non-significant** | 308 (98.4%) | 20 (10.4%) | 0.89 (0.85-0.93) |
| **Significant** | 5 (1.6%) | 172 (89.6%) | |
| ***Between 50% and 100% of OIS achieved*** | | | |
| **Non-significant** | 79 (100%) | 9 (11.5%) | 0.89 (0.81-0.96) |
| **Significant** | 0 (0%) | 69 (88.5%) | |
| ***OIS surpassed*** | | | |
| **Non-significant** | 117 (98.3%) | 18 (12.5%) | 0.85 (0.79-0.91) |
| **Significant** | 2 (1.7%) | 126 (87.5%) | |

*The percentages are calculated within the 'not significant' and 'significant' meta-analysis strata. Constant continuity correction was used for handling all zero-event arms.
CI: Confidence interval

**Table 4** Subgroup analysis by achieved levels of information size. Number and percentage of meta-analyses where the DerSimonian-Laird (DL) estimator compared to the Sidik-Jonkman (SJ) estimator yielded the same or opposite inference with regard to statistical significance.

# Figures

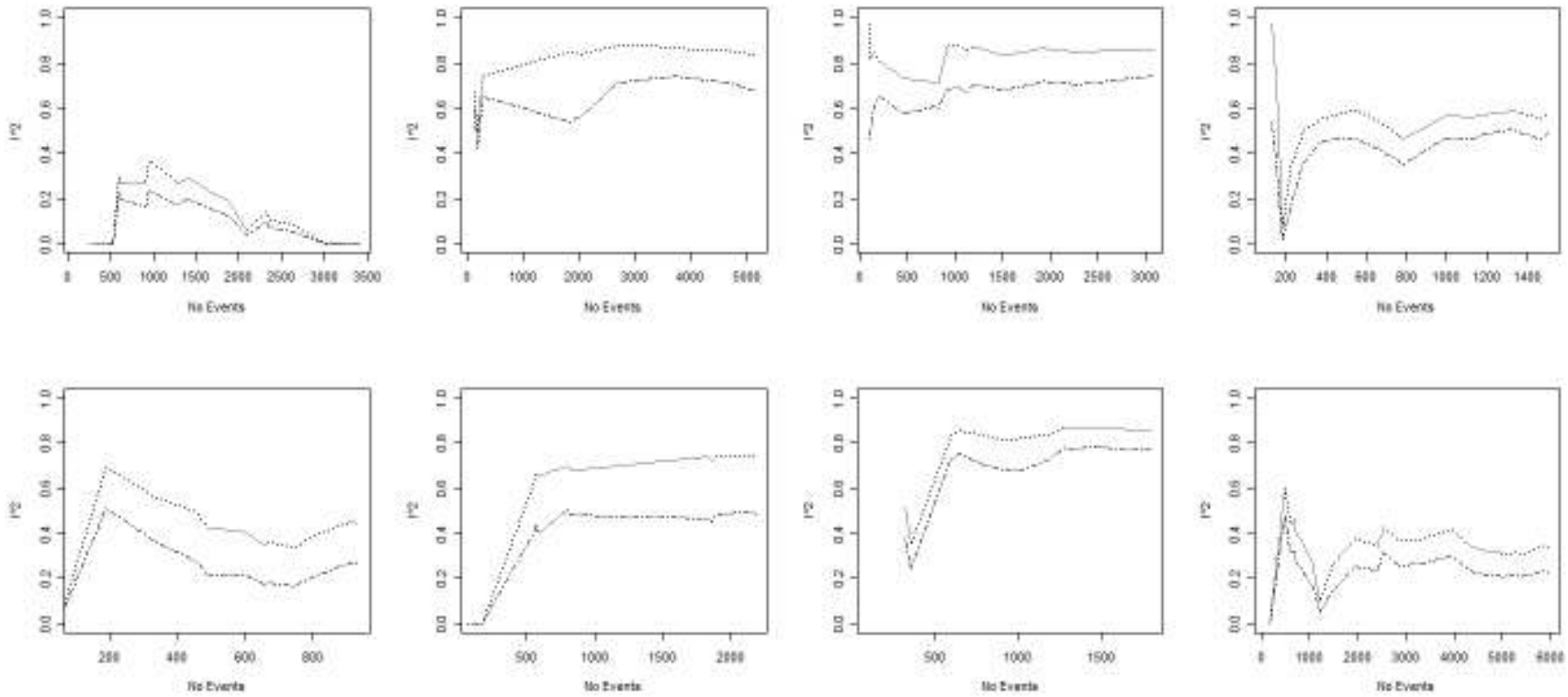



**Figure 1** Presents the evolution of the cumulative $I^2$ estimates and cumulative $D_{DL}^2$ estimates in meta-analyses (1) to (8) from chapter 3. The cumulative $I^2$ are represented by the dot-dashed line (– · –) and the cumulative $D_{DL}^2$ by the dotted line ( ·····). The cumulative heterogeneity estimates are plotted in relation to the cumulative number of events.

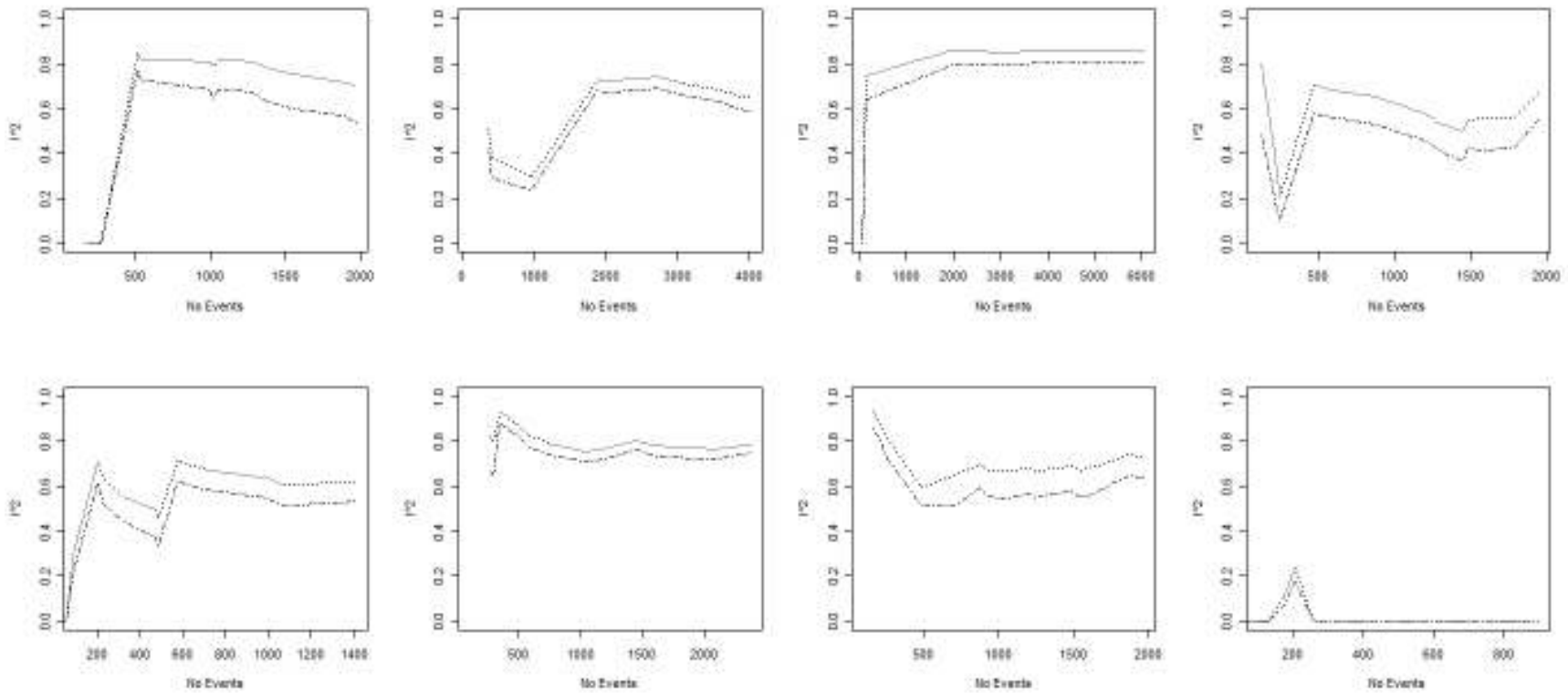



**Figure 2** Presents the evolution of the cumulative $I^2$ estimates and cumulative $D_{DL}{}^2$ estimates in meta-analyses (9) to (16) from chapter 3. The cumulative $I^2$ are represented by the dot-dashed line ( – · – ) and the cumulative $D_{DL}{}^2$ by the dotted line ( ····· ). The cumulative heterogeneity estimates are plotted in relation to the cumulative number of events.

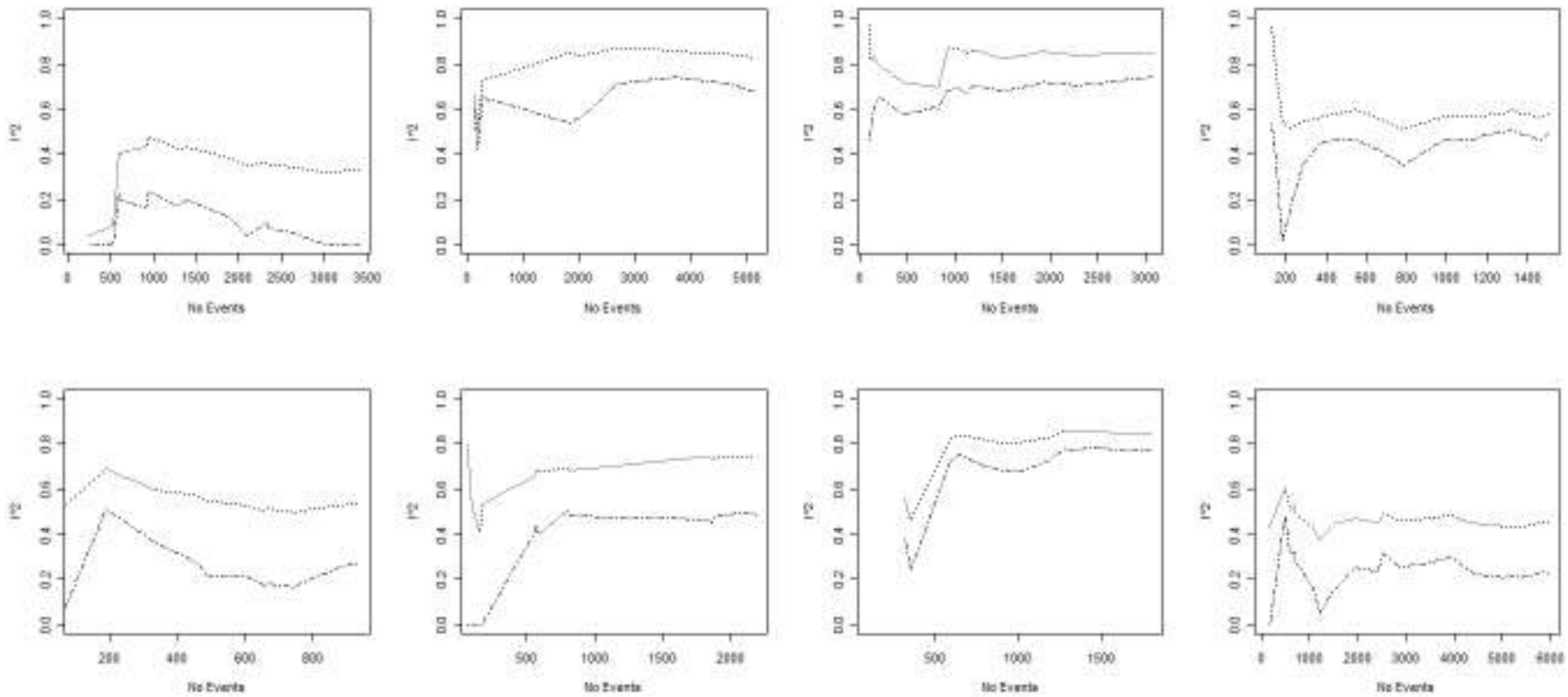



**Figure 3** Presents the evolution of the cumulative $I^2$ estimates and cumulative $D_{HM}^2$ estimates in meta-analyses (1) to (8) from chapter 3. The cumulative $I^2$ are represented by the dot-dashed line (– · –) and the cumulative $D_{HM}^2$ by the dotted line ( ·····). The cumulative heterogeneity estimates are plotted in relation to the cumulative number of events.

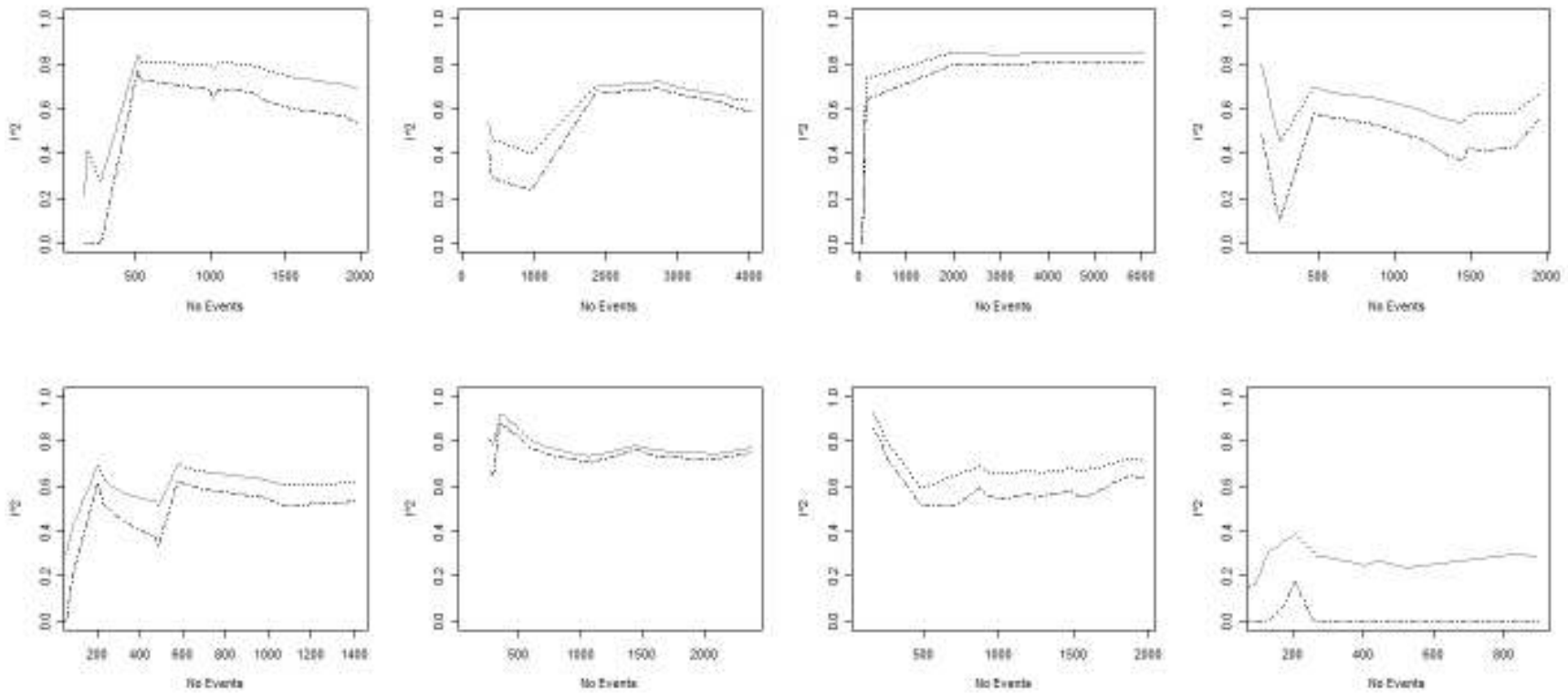



**Figure 4** Presents the evolution of the cumulative $I^2$ estimates and cumulative $D_{HM}^2$ estimates in meta-analyses (9) to (16) from chapter 3. The cumulative $I^2$ are represented by the dot-dashed line ( – · – ) and the cumulative $D_{HM}^2$ by the dotted line ( ····· ). The cumulative heterogeneity estimates are plotted in relation to the cumulative number of events.

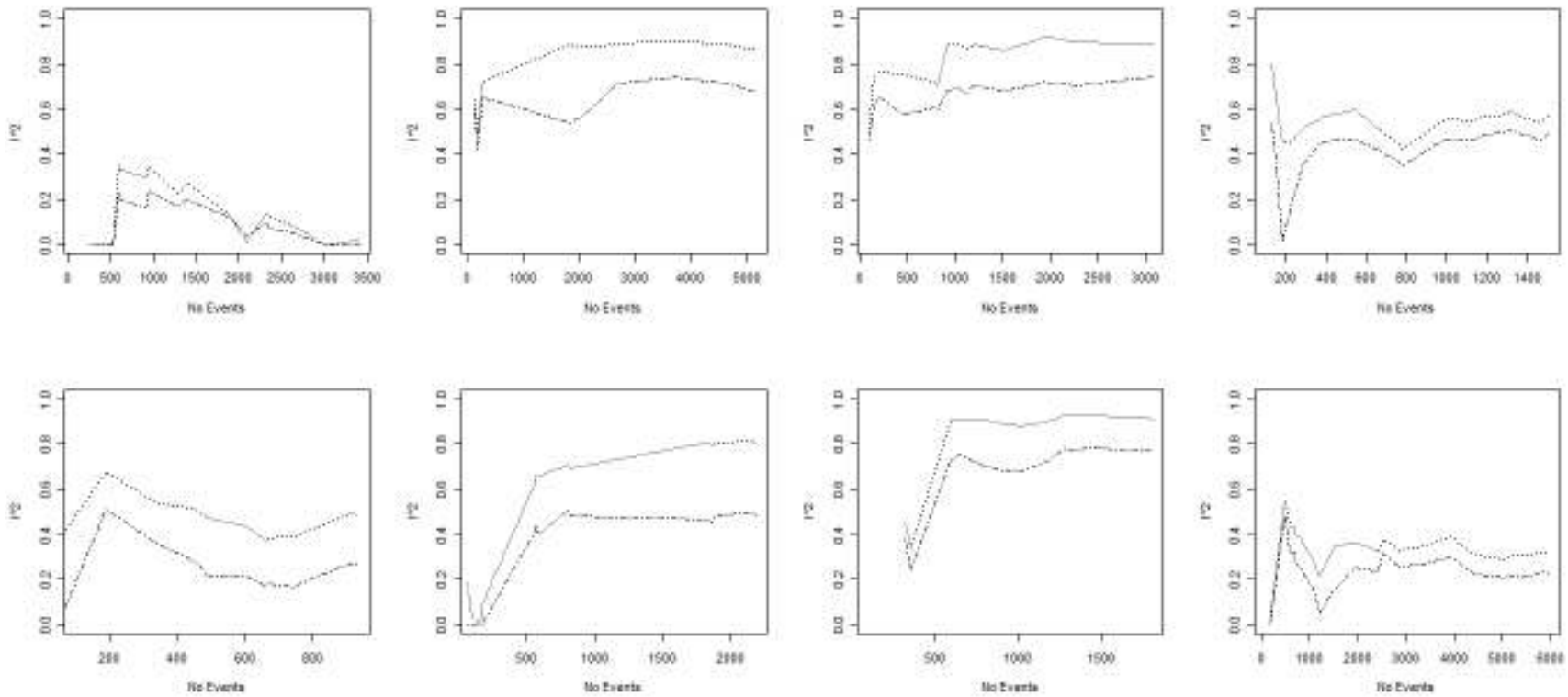



**Figure 5** Presents the evolution of the cumulative $I^2$ estimates and cumulative $D_{REML}{}^2$ estimates in meta-analyses (1) to (8) from chapter 3. The cumulative $I^2$ are represented by the dot-dashed line (– · –) and the cumulative $D_{REML}{}^2$ by the dotted line (·····). The cumulative heterogeneity estimates are plotted in relation to the cumulative number of events.

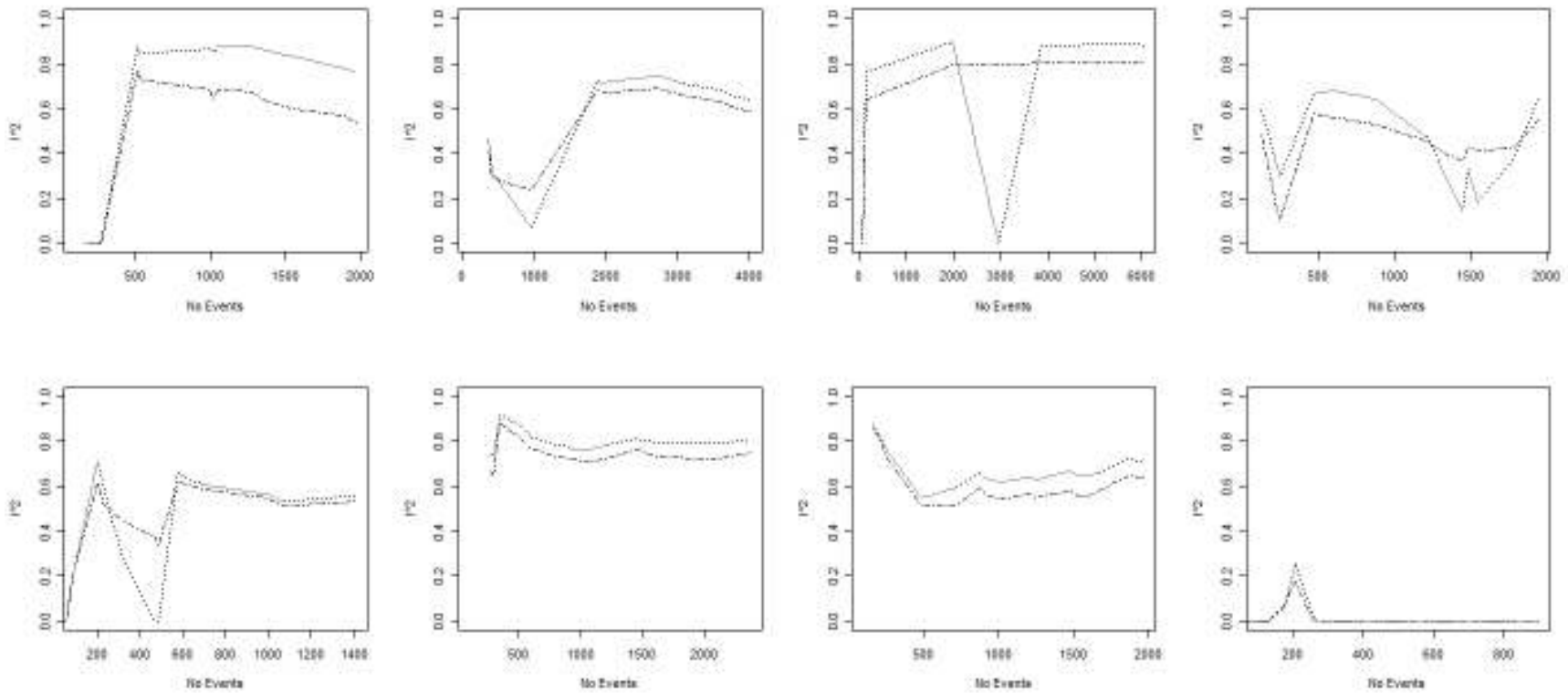



**Figure 6** Presents the evolution of the cumulative $I^2$ estimates and cumulative $D_{REML}{}^2$ estimates in meta-analyses (9) to (16) from chapter 3. The cumulative $I^2$ are represented by the dot-dashed line ( – · – ) and the cumulative $D_{REML}{}^2$ by the dotted line ( ····· ). The cumulative heterogeneity estimates are plotted in relation to the cumulative number of events.

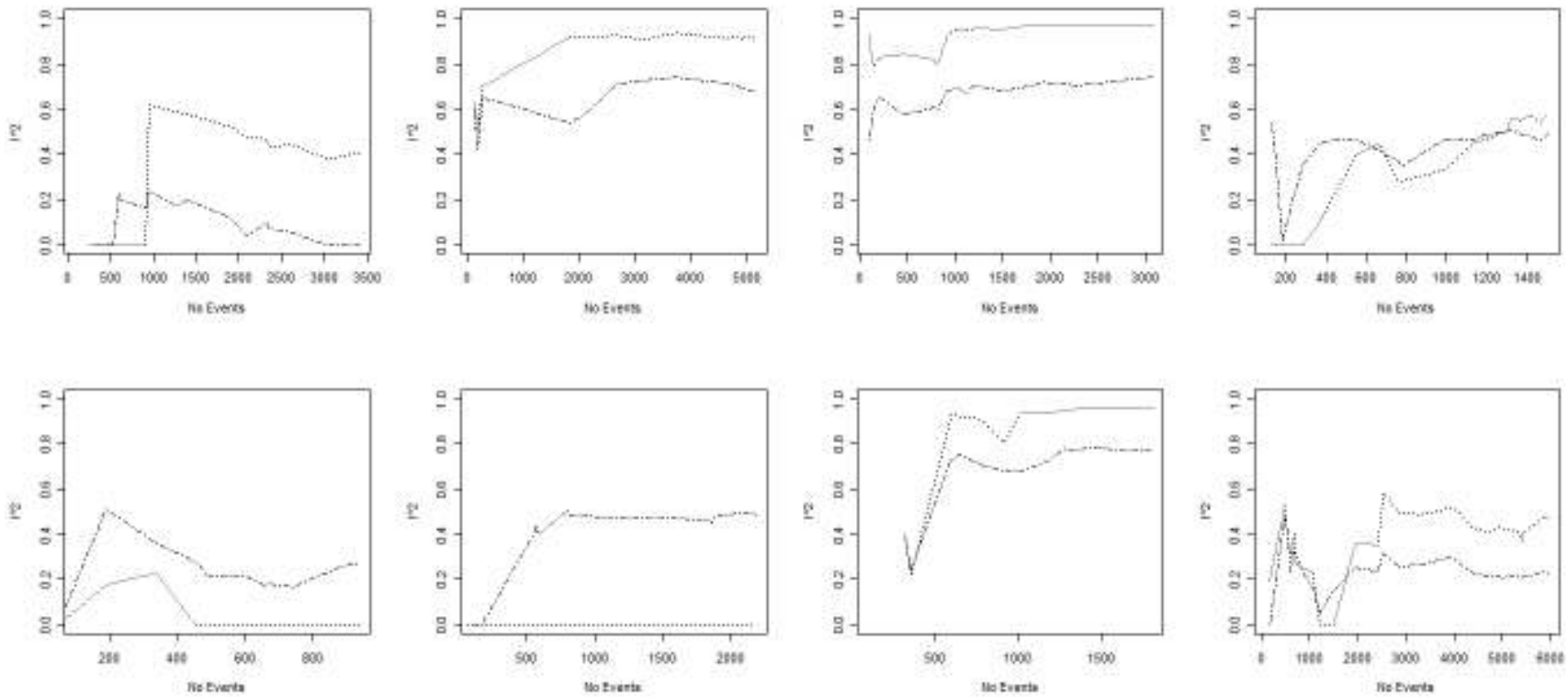



**Figure 7** Presents the evolution of the cumulative $I^2$ estimates and cumulative $D_{HE}^2$ estimates in meta-analyses (1) to (8) from chapter 3. The cumulative $I^2$ are represented by the dot-dashed line ( − · − ) and the cumulative $D_{HE}^2$ by the dotted line ( ····· ). The cumulative heterogeneity estimates are plotted in relation to the cumulative number of events.

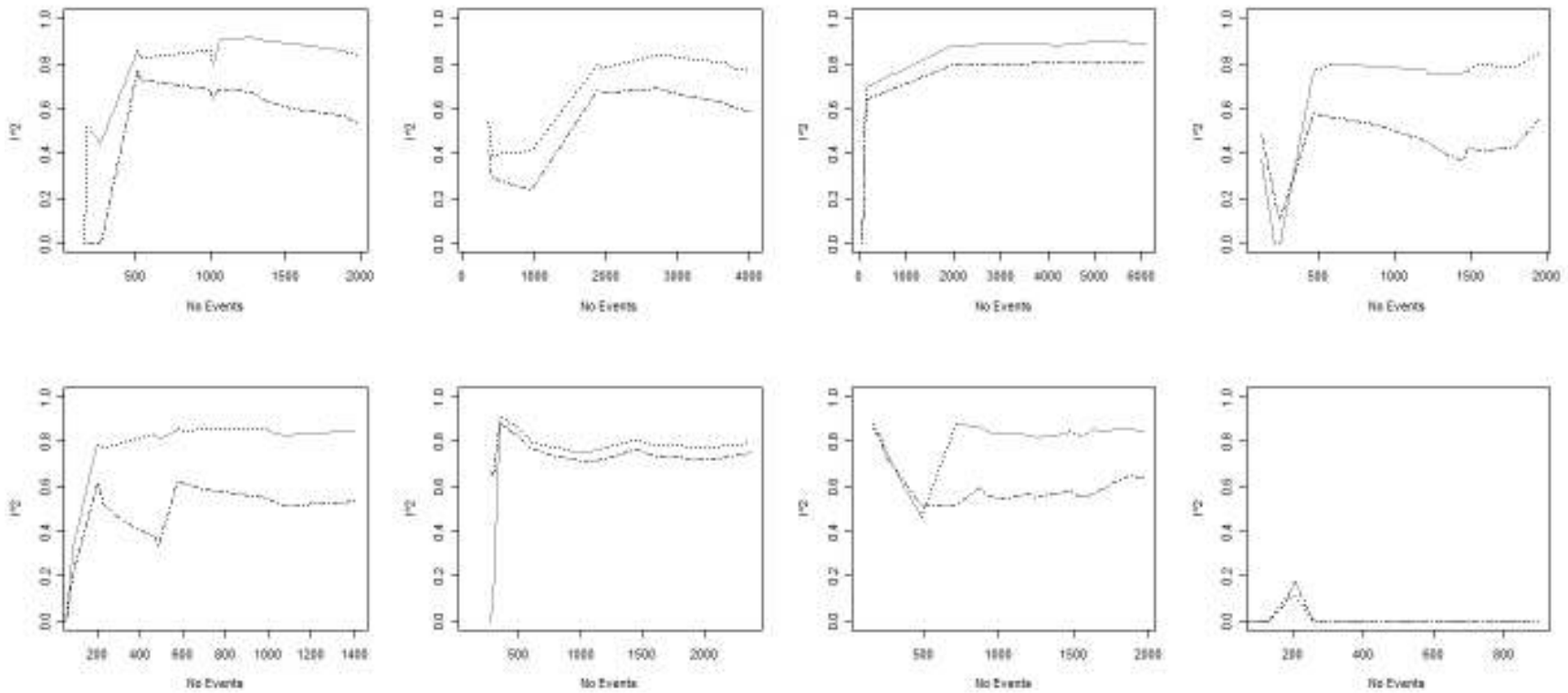



**Figure 8** Presents the evolution of the cumulative $I^2$ estimates and cumulative $D_{HE}^2$ estimates in meta-analyses (9) to (16) from chapter 3. The cumulative $I^2$ are represented by the dot-dashed line ( **– · –**) and the cumulative $D_{HE}^2$ by the dotted line ( ·····). The cumulative heterogeneity estimates are plotted in relation to the cumulative number of events.

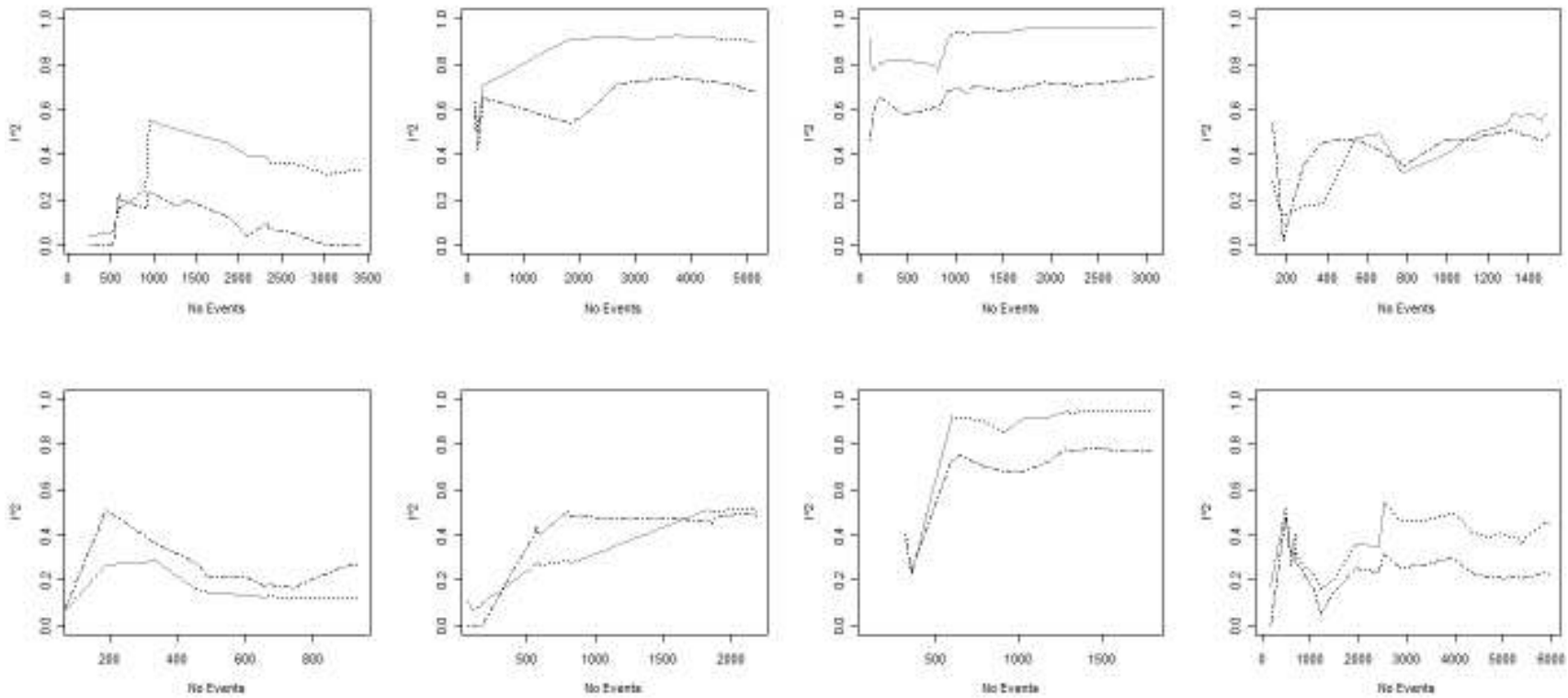



**Figure 9** Presents the evolution of the cumulative $I^2$ estimates and cumulative $D_{SJ}^2$ estimates in meta-analyses (1) to (8) from chapter 3. The cumulative $I^2$ are represented by the dot-dashed line (– · –) and the cumulative $D_{SJ}^2$ by the dotted line ( ····· ). The cumulative heterogeneity estimates are plotted in relation to the cumulative number of events.

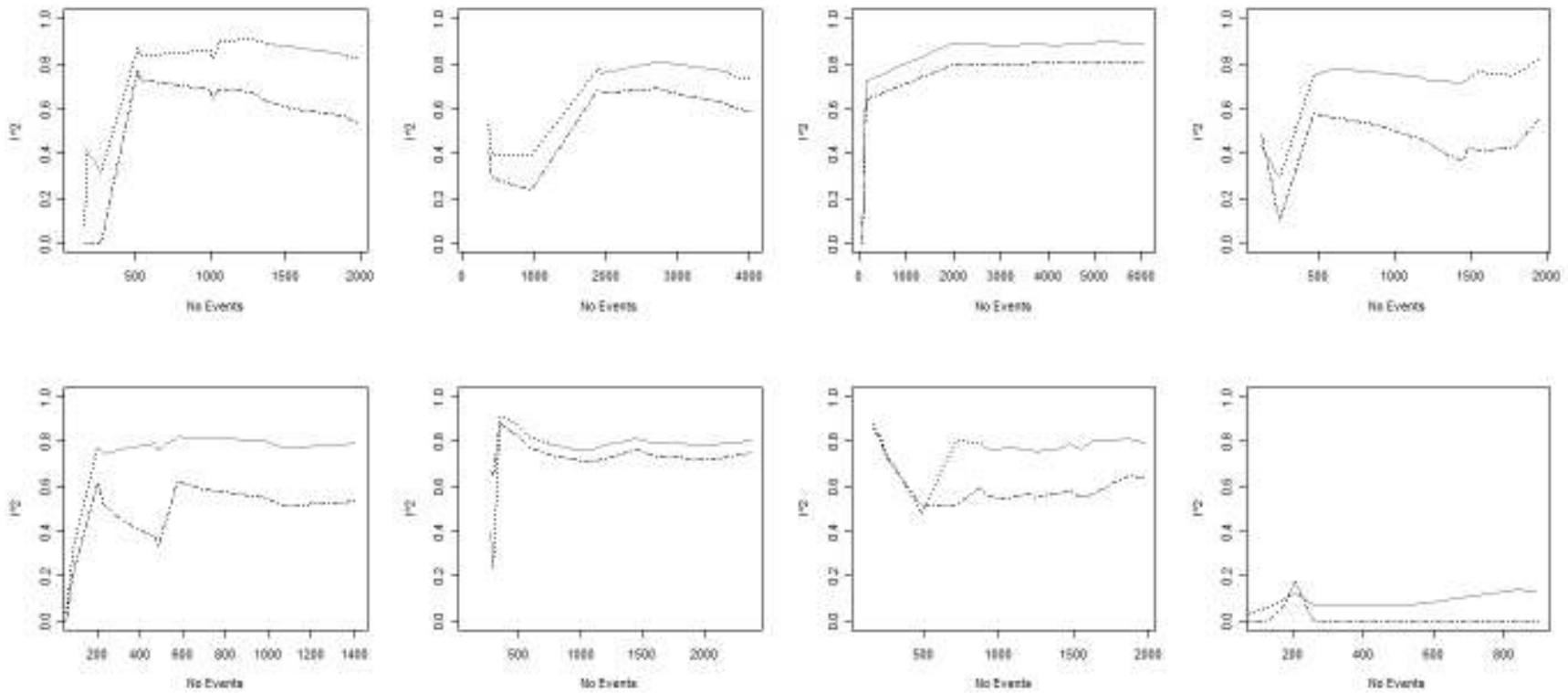



**Figure 10** Presents the evolution of the cumulative $I^2$ estimates and cumulative $D_{SJ}^2$ estimates in meta-analyses (9) to (16) from chapter 3. The cumulative $I^2$ are represented by the dot-dashed line (– · –) and the cumulative $D_{SJ}^2$ by the dotted line (·····). The cumulative heterogeneity estimates are plotted in relation to the cumulative number of events.

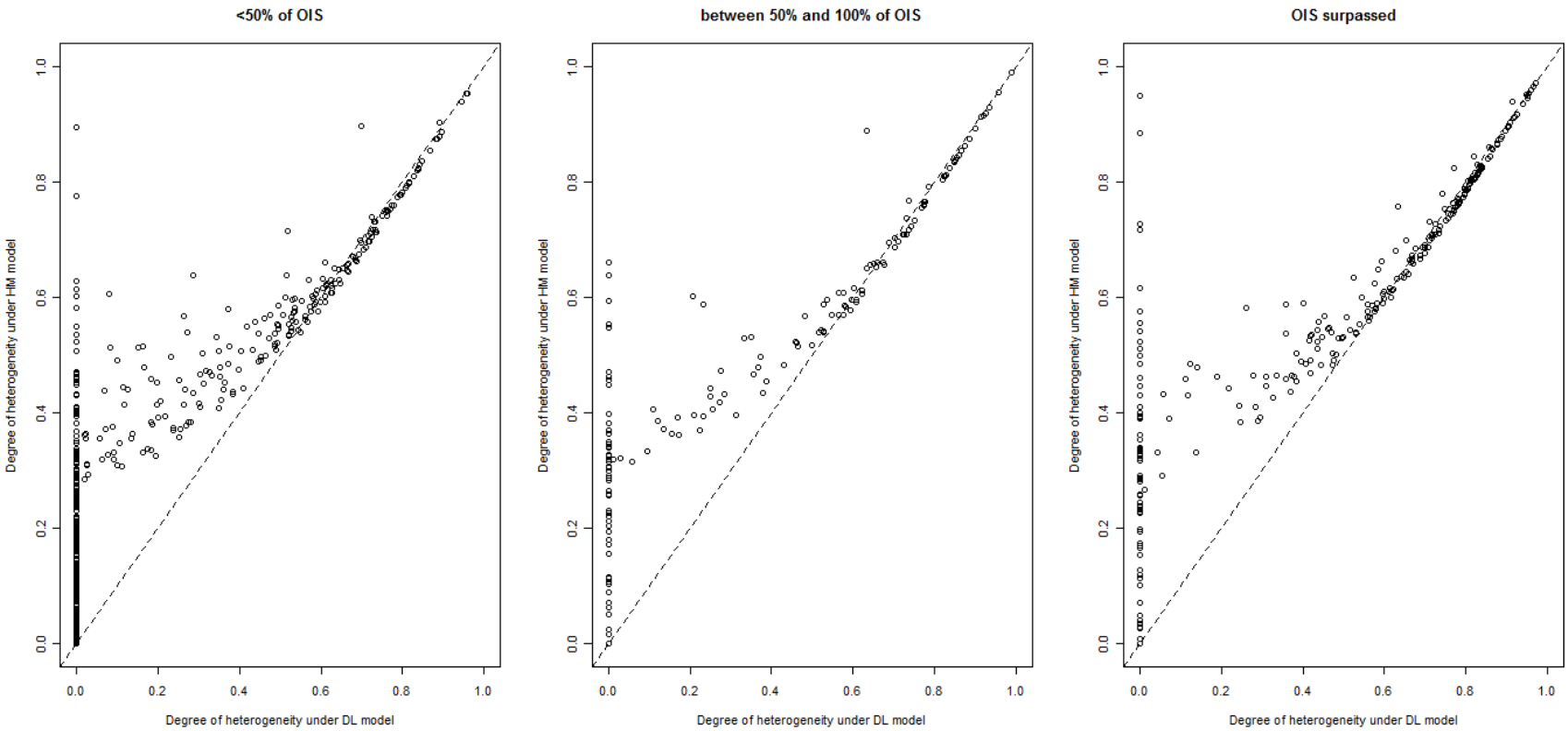



**Figure 11** Presents of degree of heterogeneity under the DerSimonian-Laird random-effects model (x-axis) plotted against the degree of heterogeneity under the Hartung-Makambi (HM) random-effects model (y-axis) for the subsets of meta-analyses where less than 50% of the OIS is achieved, between 50% and 100% of the OIS is achieved, and where the OIS is surpassed.

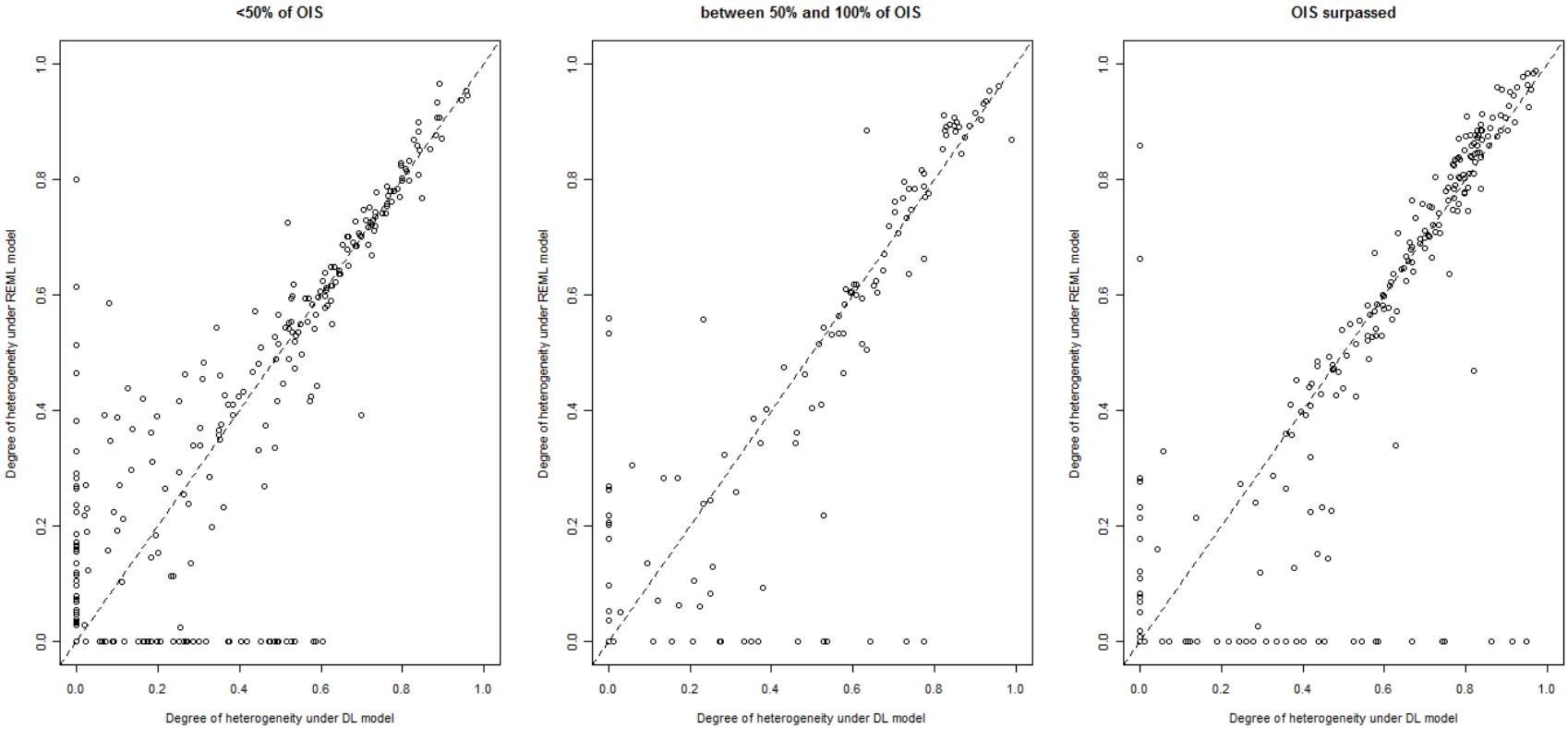



**Figure 12** Presents of degree of heterogeneity under the DerSimonian-Laird random-effects model (x-axis) plotted against the degree of heterogeneity under the restricted maximum-likelihood (REML) random-effects model (y-axis) for the subsets of meta-analyses where less than 50% of the OIS is achieved, between 50% and 100% of the OIS is achieved, and where the OIS is surpassed.

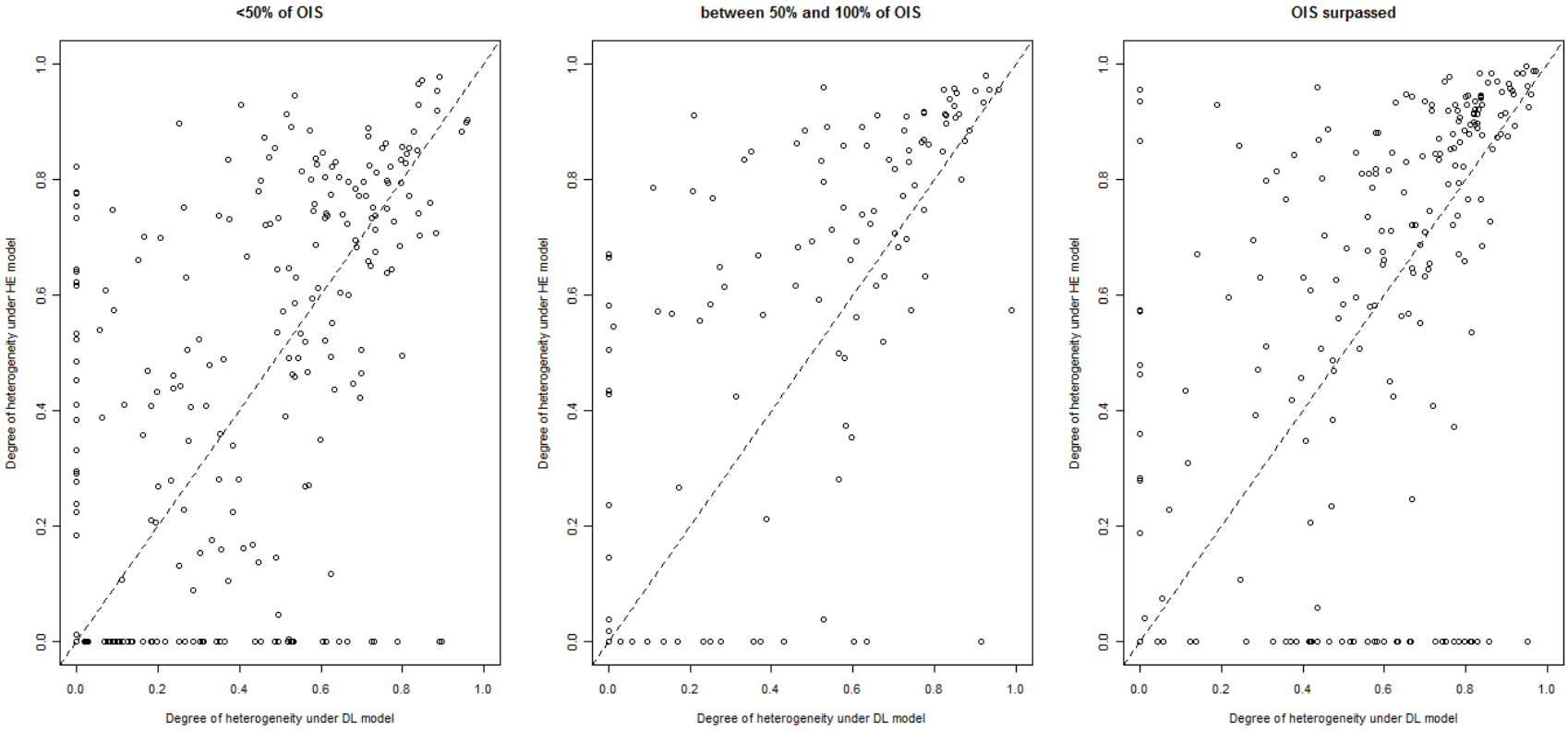



**Figure 13** Presents of degree of heterogeneity under the DerSimonian-Laird random-effects model (x-axis) plotted against the degree of heterogeneity under the Hedges (HE) random-effects model (y-axis) for the subsets of meta-analyses where less than 50% of the OIS is achieved, between 50% and 100% of the OIS is achieved, and where the OIS is surpassed.

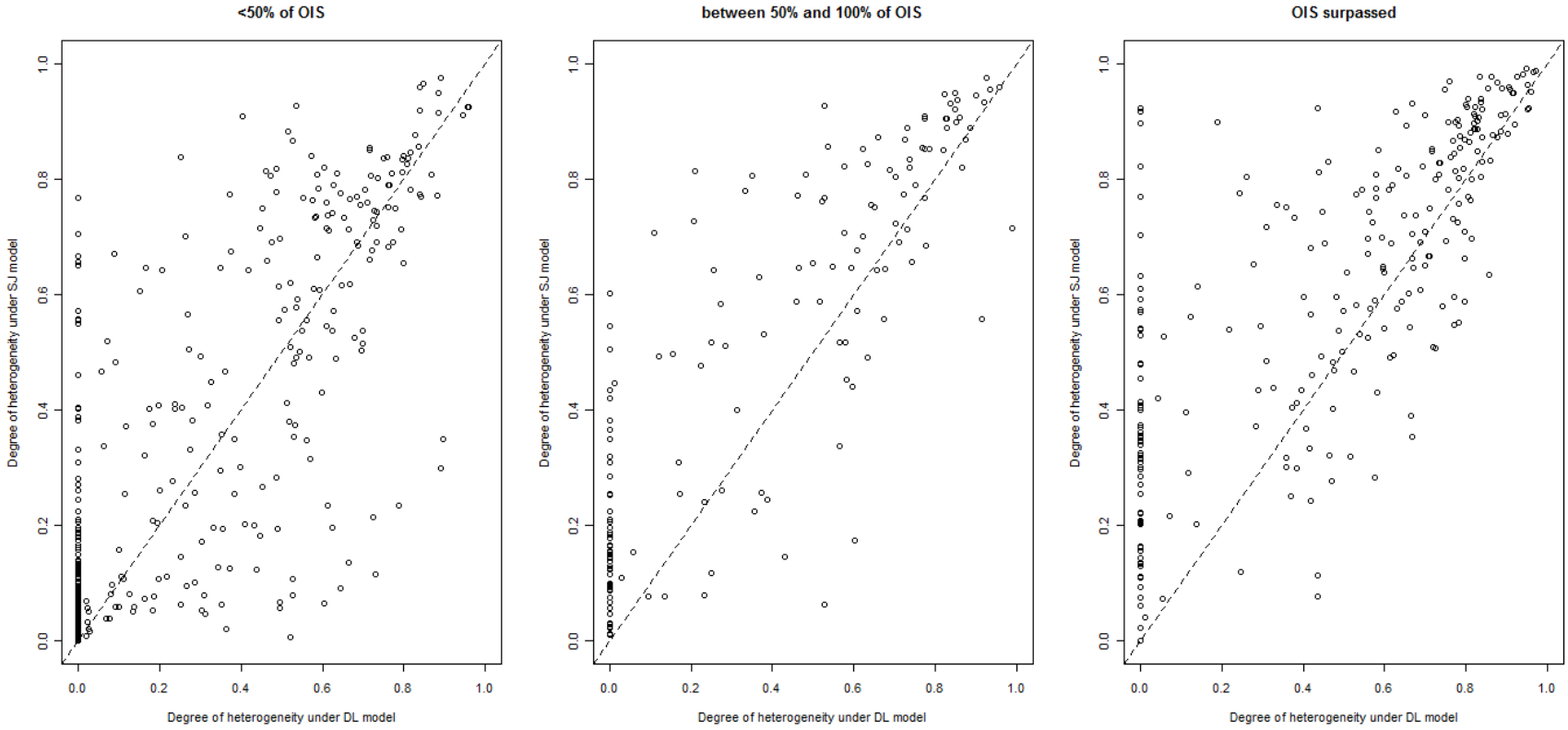



**Figure 14** Presents of degree of heterogeneity under the DerSimonian-Laird random-effects model (x-axis) plotted against the degree of heterogeneity under the Sidik-Jonkman (SJ) random-effects models (y-axis) for the subsets of meta-analyses where less than 50% of the OIS is achieved, between 50% and 100% of the OIS is achieved, and where the OIS is surpassed.

# Chapter 7: Discussion

This chapter is structured as follows. First, the key findings of the chapters 2 to 6 are summarized in brief. Second, I consider whether the studies presented in chapters 2, 3, 4, and 5 in collection meet the objectives of this thesis. Third, I reflect on how the findings might impact meta-analytic practice, either single-handedly or in concert with other past and future publications. Lastly, because proper dissemination of statistical issues in meta-analysis to a clinical audience is an overarching topic of this thesis, I reflect on the dissemination challenges incurred in the presented studies.

**Summary of findings**

The studies presented in chapter 2, 3, 4 and 5 all contribute to furthering our understanding of statistical metrics in meta-analysis that are conceptually relatively simple.

In chapter 2, the simulations demonstrated that random error (chance) plays an important role in estimation of intervention effects in meta-analysis: reaching the optimal information size (OIS, i.e., the required meta-analysis sample size) provides good protection against overestimation due to random error. On the more complex side, the simulation findings provide insights on the magnitude and risk of overestimating an intervention effect due to random error before surpassing the OIS.

In chapter 3 it was shown empirically that the reliability of the popular measure of heterogeneity, $I^2$, does too depend on the amount of accumulated evidence. Theoretical arguments and empirical evidence were provided to demonstrate that $I^2$ can and often will incur considerable fluctuations as evidence accumulates and that these fluctuations occur when the evidence is

sparse (e.g., when the meta-analysis includes less than 200 events). Further, the 95% confidence intervals for $I^2$ demonstrated satisfactory performance.

In chapter 4 three popular meta-analytic inferential measures (the p-value and the 95% confidence intervals for the meta-analysed intervention effect, and the degree/percentage of heterogeneity) from the conventional DerSimonian-Laird random-effects were compared with the same three inferential measures from four selected alternative random-effects models. First, it was shown empirically that while p-values and 95% confidence intervals based on alternative random-effects models do not infrequently differ from the ones of the DerSimonian-Laird random-effects model, such differences are rarely of a magnitude that would cause alterations in the statistical inferences about the overall intervention effect, let alone the conclusion of the systematic review in which the random-effects meta-analysis is included. Second, it was shown that the estimated degree (percentage) of heterogeneity in meta-analyses based on alternative random-effects models often differ substantially from that of the DerSimonian-Laird random-effects model. One of the four alternative random-effects models - the Hartung-Makambi model - yielded heterogeneity estimates that behaved in a stable and predictable manner in relation to the DerSimonian-Laird model. The remaining three models did not.

In chapter 5 the strengths and limitations of twelve methods for enhancing interpretability of continuous outcome meta-analyses were reviewed. Their performance was assessed in three illustrative examples, and recommendations for which of the reviewed methods to use were provided in the form of a simple 2-stage algorithm.

In chapter 6 two additional analyses cast light on issues in chapter 2, 3 and 4 that were interrelated. The first additional analysis explored the extent to which the measured degree of heterogeneity based on different random-effects models fluctuated over time. It was shown that

the measure based on the Hartung-Makambi model fluctuated less and became stable earlier on when compared with the measure based on the DerSimonian-Laird model (both $I^2$ and $D_{DL}^2$). Measuring the degree of heterogeneity with the three other alternative random-effects models, however, resulted in larger fluctuations and required more evidence to become stable. The second analysis explored whether the intensity of inferential discrepancies observed in chapter 4 would vary with the meta-analysis information size. Only minimal changes were observed.

**Thesis objectives – were they met?**

To recapitulate, the objectives of this thesis were to fill some of the demand for statistical contributions to meta-analysis where

1) improvements to current statistical practice in meta-analysis are conveyed at the level that most systematic review authors will be able to understand

2) current statistical methods which are widely applied in meta-analytic practice undergo thorough testing and examination.

The key finding in chapter 2 - that the optimal information size provides good protection against overestimation of intervention effects – signifies an important milestone in the effort to promote information size requirements in meta-analysis. The finding confirms the theoretical arguments supporting information size requirements, and it will provide proponents of the OIS with much better basis for promoting the approach. The second key finding - that random error does pose a problem before reaching the OIS – is sufficiently simple and strong to meet objective 1. Hopefully this finding will aid in instilling the necessary additional caution among authors when it comes to interpretation of intervention effects. The many figures in chapter 2 from which

authors of meta-analyses can determine the approximate risk of overestimation due to random error are extremely comprehensive. It is unlikely that systematic review authors will find these figures easy to use in practice. In this regard, objective 1 was not met. It should, however, be noted that this manuscript originally only included the simulations based on the survey of Cochrane Heart group mortality meta-analyses (thus, a third of what it does now), but was expanded as a result of peer reviewer comments. In its original format, the tables and figures were much less comprehensive, and would likely have been interpretable to systematic review authors in cardiology.

In chapter 3 the theoretical properties of $I^2$ and its 95% confidence interval were described in an as non-statistical as possible language. The 16 plots and the reported span of fluctuations are easy to interpret. In this capacity, the study presented in chapter 3 meets thesis objective 1. The demonstrated substantial fluctuations of $I^2$ and the satisfactory performance of the associated 95% confidence intervals are scientifically important. Properties of the $I^2$ measure and its associated 95% confidence intervals have not previously been studied empirically in a temporal framework. In this capacity, the study presented in chapter also meets thesis objective 2.

In chapter 4, thesis objectives 1 and 2 were only partially met. The potential impact of using different random-effects models were described in fairly lay language in the introduction. This was further elaborated on in the description of the measures of inferential discrepancies in the methods section. Further, the discussion also offers some guidance for meta-analytic practice given the findings of this study. The descriptions of the random-effects methods, however, were written in a statistically heavy language. Earlier versions of the manuscript included only conceptual descriptions, but due to reviewer requests for more statistical detail, this was changed. Objective 2 was met with regards to inferences about the intervention effect, but not with regards

to inferences about the degree of heterogeneity. It was evident the choice of random-effects model does not have an important impact on the inferences about the overall intervention effect, let alone the conclusion of the systematic review that includes the meta-analysis. Thus, this finding is comforting since the contrary would have indicated that many statistically significant findings in published DerSimonian-Laird random-effects meta-analyses would likely not have been statistically significant had other random-effects models been used. The study also helps to direct research efforts in meta-analysis since clearly further simulations studies exploring inferences about the overall intervention effect would not be worthwhile. The comparisons of the degree of heterogeneity from the considered random-effects models does not directly apply to objective 2 since none of the considered heterogeneity measures are applied widely in practice. However, $D_{DL}^2$ is a reasonable surrogate for $I^2$ in this capacity and so, objective 2 is met indirectly.

Chapter 5, although fairly statistical, meets objective 1 for a number of reasons. First, the strengths and limitations of the reviewed methods have never before been described collectively. Second, most of the reviewed methods were presented in a highly statistically oriented format in reviewed methodological papers. Lastly, and probably most importantly, the provided recommendations are tailored to cater to a clinical audience, and they are extremely simple, so most systematic review authors with basic biostatistics training should be able to follow them. Objective 2 was also met. Most importantly, many of the properties in Table 1 have not previously been mentioned in the literature. The performances of the reviewed methods were also contrasted in Table 2.

**How are the findings likely to impact meta-analytic practice?**

In this section I discuss how each of the chapters are likely to impact meta-analytic practice singlehandedly or in collection with other research or initiatives. For this discussion it is assumed that the manuscripts presented in chapter 2, 3, 4, and 5 will be published in peer reviewed journals in the near future.

Chapter 2 will probably not have much impact on the rate with which OIS becomes an increasingly integral part of meta-analysis. The GRADE working group is already doing their part to ensure widespread use of OIS. Further, software for calculating the OIS is expected to be released in August 2011 and may aid the propagation of OIS.[1] Authors of meta-analyses also need to pay attention to random error when the OIS is not reached, and analyse and interpret their data accordingly. Again, the GRADE working group provides recommendations for such conduct. Other leading groups of methodologists, for example, the Cochrane Collaboration Statistical Methods Group are now debating this issue (e.g., via their mailing list). The issue of random error in meta-analysis has also received a great deal of attention in the methodological literature over the past five years.[2-14]

The immediate impact of the study presented in chapter 3 is unlikely to be profound. The study is not the first to examine various properties of $I^2$ and its associated 95% confidence intervals, and it is not the first to describe various properties of $I^2$ and its associated 95% confidence intervals in a lay language. Other studies have even been of a grander scale. It thus seems that methodological publications on this topic singlehandedly do not suffice to change meta-analytic practice. In collection with past and future studies, this study may contribute sufficiently to the awareness of the need for reporting 95% confidence intervals for $I^2$. Once sufficient awareness has been created, it is likely that 95% confidence intervals for $I^2$ will become standard in Review

Manager and other meta-analysis software packages. To create sufficient awareness about the risk of fluctuations of $I^2$, it is likely that specific meta-epidemiologic evidence about the causes of fluctuations will need to emerge. By analogy, if no evidence existed about the sources of bias on intervention effects, it is likely that meta-analysts would be less cautious about interpreting intervention effects.

Chapter 4 leaves unanswered questions with regard to the estimation of the degree (percentage) of heterogeneity. In the first of the additional analyses in chapter 6, the degree of heterogeneity based on the Hartung-Makambi estimator seems preferable. However, 95% confidence intervals have not been developed for the $D^2$ type of heterogeneity measure. If this proves to be too challenging of a task, one could instead derive an $I^2$ type heterogeneity measure based on the Hartung-Makambi model and then derive 95% confidence intervals in a similar fashion as done by Higgins and Thompson.[15] In any case, the road to widespread dissemination would be long. It would require mathematical derivations, comprehensive simulation studies, empirical confirmatory studies, and lastly, didactic papers and other efforts by groups of methodologists.

Chapter 5, in time, will likely have great impact on meta-analytic practice. It is the first in a series of related papers and initiatives. It is deliberately methodologically exhaustive since the authors of this paper required in-depth knowledge of the reviewed methods to write up two didactic manuscripts post its publication. One subsequent manuscript will be a light version which is targeted even more to a clinical audience. This paper is also the 13$^{th}$ in a series of GRADE papers planned to be published in the *Journal of Clinical Epidemiology*. The second subsequent work that will be undertaken as a result of this paper is a chapter in the Cochrane Handbook on reporting and analysing health related quality of life and patient reported

outcomes. In addition, empirical studies and methodological extensions are already underway.[16,17]

**Further reflections**

An overarching topic in this thesis is the need for methodological papers that can properly describe statistical issues in meta-analysis to a clinical audience. In this thesis, this has been done through simulation, empirical study, and methodological review. Each approach requires an in-depth statistical understanding of the issues at play before one can even begin to properly re-articulate these issues in a language that will resonate well with clinicians. For statisticians, however, once one possesses such in-depth knowledge many other complex statistical questions become apparent. Many statisticians may find it more compelling to dedicate their time to answering such complex questions rather than dedicating their time to verifying simple implicit assumptions (e.g., that $I^2$ is reliable over time) and communicating these issues at a basic statistical level. Another challenge with increasing the number of scientific publications that meet objective 1 or 2 of this thesis is the peer review system that journals use. When submitting manuscripts that fall into the category of the manuscripts in this thesis, the journal editors will typically assign at least one statistical expert as a peer reviewer. This person will likely request more in-detail explanations of the statistical issues (including more complex equations), and will not take into account what communication style is most likely to resonate well with clinicians. This, for example, was the case with the manuscripts presented in chapter 2 and 4. Nevertheless, there is reason to believe better times are ahead. The number of online access journals have soared this past decade and with it has the number of specialized journals. One journal, Research Synthesis Methodology, has emerged to deal with methodological issues in meta-analysis only,

and one journal, BMC Systematic Reviews, has emerged to deal with methodological issues in systematic reviews only. Perhaps such journals will increasingly put emphasis on what is needed to facilitate the next natural advances in practice.

**Concluding remarks**

In conclusion, this PhD thesis has explored issues of estimation and interpretation of meta-analysed intervention effects and heterogeneity. The manuscripts in this thesis were designed to deal with issues closely related to current widespread meta-analytic practice. Each manuscript makes contributions by either eliminating issues of uncertainty about the credibility of methods that are or will soon be widely used, or by demonstrating problems with current methodological issues and subsequently educating on and demonstrating the superiority of alternative relatively simple methods. Overall, all of the manuscripts presented in the thesis met at least one of the two thesis objectives, and each manuscript constitutes an important contribution to the advance of meta-analytic practice.

## References


(1) TSA v.0.8 - Trial Sequential Analysis software application. 2011. http://www.ctu.dk/tsa

(2) Bollen C, Uiterwaal C, Vught A, Tweel Ingebo. Sequential meta-analysis of past clinical trials to determine the use of a new trial. *Epidemiology* 2006; 17:644-649.

(3) Borm GF, Donders AR. Updating meta-analyses leads to larger type I errors than publication bias. *J Clin Epidemiol* 2009; 62(8):825-830.

(4) Brok J, Thorlund K, Gluud C, Wetterslev J. Trial sequential analysis reveals insufficient information size and potentially false positive results in many meta-analyses. *Journal of Clinical Epidemiology* 2008; 61:763-769.

(5) Brok J, Thorlund K, Wetterslev J, Gluud C. Apparently conclusive meta-analyses may be inconclusive - Trial sequential analysis adjustment of random error risk due to repetitive testing of accumulating data in apparently conclusive neonatal meta-analyses. *International Journal of Epidemiology* 2009; 38:287-298.

(6) Higgins JPT, Whitehead A, Simmonds M. Sequential methods for random-effects meta-analysis. *Stat Med* 2011; 30(9):903-921.

(7) Hu M, Cappeleri J, LanKK. Applying the law of the iterated logarithm to control type I error in cumulative meta-analysis of binary outcomes. *Clinical Trials* 2007; 4:329-340.

(8) Lan KK, Hu M, Cappelieri J. Applying the law of the iterated logarithm to cumulative meta-analysis of a continuous endpoint. *Statistica Sinica* 2003; 13:1135-1145.

(9) Thorlund K, Devereaux PJ, Wetterslev J, Guyatt G, Ioannidis JP, Thabane L et al. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? *International Journal of Epidemiology* 2009; 38:276-286.

(10) Thorlund K, Anema A, Mills E. Interpreting meta-analysis according to the adequacy of sample size. An example using isoniazid chemoprophylaxis for tuberculosis in purified protein derivative negative HIV-infected individuals. *Clinical Epidemiology* 2010; 2:57-66.

(11) Trikalinos TA, Churchill R, Ferri M, Leucht S, Tuunainen A, Wahlbeck K et al. Effect sizes in cumulative meta-analyses of mental health randomized trials evolved over time. *Journal of Clinical Epidemiology* 2004; 57(11):1124-1130.

(12) van der Tweel I, Bollen C. Sequential meta-analysis: an efficient decision making tool. *Clinical Trials* 2010; 7(2):136-146.

(13) Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. *Journal of Clinical Epidemiology* 2008; 61:64-75.

(14) Wetterslev J, Thorlund K, Brok J, Gluud C. Estimating required information size by quantifying diversity in a random-effects meta-analysis. *BMC Medical Research Methodology* 2009; 9(86).

(15) Higgins JPT, Thompson S. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 2002; 21:1539-1558.

(16) da Costa BR, Rutjes AWS, Johnston BC, ReichenBach S, Nuesch E, Guyatt GH et al. Standardized mean differences may be used to derive odds ratios of treatment response: meta-epidemiological study (abstract). Cochrane Colloquium 2011 abstract book . 2011.

(17) Johnston BC, Thorlund K, da Costa BR, Furukawa TA, Guyatt GH. Improving the Interpretation of Meta-analyses Using Minimal Important Difference: Combining Scores from Continuous Outcome Measures (abstract). 19thth Cochrane Coloquium abstract book . 2011.