

UNIVERSITY OF COPENHAGEN
FACULTY OF HEALTH AND MEDICAL SCIENCES
&
MACQUARIE UNIVERSITY
FACULTY OF SCIENCE AND ENGINEERING



MACQUARIE
University
SYDNEY · AUSTRALIA



PhD thesis

Statistical methods for meta-analysis

Anne Lyngholm Soerensen

Academic advisors: Ian C. Marschner, Theis Lange & Christian Gluud

This thesis has been jointly submitted to:
The Graduate School of the Faculty of Health and Medical Sciences, University of Copenhagen, and,
The School of Mathematical and Physical Sciences of the Faculty of Science and Engineering, Macquarie University
on March 22, 2024

Preface

This thesis has been submitted to the Graduate School of Health and Medical Sciences at University of Copenhagen, Denmark, and to the School of Mathematical and Physical Sciences at Macquarie University, Sydney, Australia. The work for this thesis was carried out part-time at the Section of Biostatistics, Department of Public Health, University of Copenhagen and part-time at the Department of Mathematics and Statistics, Macquarie University. This PhD was partly funded by the International Cotutelle Macquarie University Research Excellence Scholarship (Cotutelle “iMQRES”) and Copenhagen Trial Unit, Rigshospitalet, Denmark.

Cotutelle

This is a Cotutelle PhD thesis. Being a Cotutelle PhD, I was enrolled at two universities with each their principal supervisor at the same time. The two universities were Macquarie University in Sydney, Australia, and University of Copenhagen in Denmark. Choosing a Cotutelle PhD increased the supervisor team, which turned out great as I found myself to be spoiled by having many great supervisors. Without the supervisor team, I would not have made it through this PhD which included 1 year of conversations with legal teams creating the contract, Australian visa problems, bush fires in New South Wales, the COVID-19 pandemic, lock downs in two countries and a bub, aka baby.

Thank you to Ian C. Marschner for being a consistent and encouraging supervisor. I think everyone will have a successful and good experience doing a PhD if they have you as a supervisor. Thank you for introducing me to Sydney. Unfortunately I now feel very attached to Sydney and have to return at some point. I partly blame you for this.

Thank you to Theis Lange and Christian Gluud for being very supporting supervisors. Especially for letting me try to find my own research ideas while helping me along the way. I feel like I have grown more confident in being an independent researcher with you. Without Christian and the Copenhagen Trial Unit this PhD would not have been financially possible and I am really grateful for the trust.

Thank you to Markus Harboe Olsen for helping with the RTSA package. How lucky I am to have had your help.

And at last, thank you to my family and friends, especially my husband Oliver for sticking with me no matter the country and to my son Frej who keeps me on my toes.

Abstract

A meta-analysis is a statistical analysis method which synthesises the results of multiple trials into a single result. Rather than focusing on one of the individual trials' conclusions, a meta-analysis can describe in a concise manner the body of evidence. If the meta-analysis is created using studies with low risk of bias, the results of a meta-analysis are considered to be the highest standard of evidence. However, the quality of the meta-analysis is also dependent on the statistical method used for modelling the data. Many factors can lead to bias, invalidity or misinterpretation of the results of the meta-analysis. This thesis presents a statistical investigation of two such areas: subgroup meta-analysis and sequential meta-analysis.

The first area, subgroup meta-analysis, relates to ecological bias in aggregate data subgroup meta-analysis. Ecological bias is a type of bias caused by the ecological level of a subgroup which modifies the effect of treatment on outcome. Ecological bias may lead to confounding that biases the assessment of subgroup and treatment effect interactions, depending on the analysis method used. This thesis proposes a new method for subgroup meta-analysis using linear mixed models. The method is used for estimating effect modification of treatment by subgroup while correcting for ecological bias. This method fills gaps where the existing methods fall short.

The second area, sequential meta-analysis, is concerned with the validity of statistical results when a meta-analysis is updated sequentially. Updating a meta-analysis will increase the risk of finding a false-positive. If properly planned in advance using a group sequential design, the inflation of type-I-error may be removed. Trial Sequential Analysis is a software package developed in Java to adapt the group sequential designs originally created for single trials to meta-analyses. A new version of the software is presented with multiple extensions. New features include updated methods for sample size calculation, a framework for prospective meta-analysis, binding and non-binding futility boundaries for both one- or two-sided designs, and an extended library for calculating inference. A software package for the statistical software language R is presented.

A further problem considered within sequential meta-analysis is a specific type of bias. Conditional bias due to a decision to continue a meta-analysis happens when new studies are motivated by earlier information, e.g. the results of a promising but not definitive meta-analysis. If the new studies are combined with the promising meta-analysis this results in an upwards bias of the point estimate. Inspired by the adjustment estimators developed for group sequential methods for single trials, this thesis presents an estimator for updating meta-analyses adjusting for the conditional bias from decision making.

Resumé

En meta-analyse er en statistisk analysemetode som sammenfatter resultaterne fra flere forsøg til et generelt resultat. I stedet for at fokusere på alle de individuelle forsøgsresultater kan en meta-analyse kortfattet beskrive samlingen af evidens. Resultatet af en meta-analyse er ofte betragtet som en af de bedste former for evidens for en given hypotese. Da en meta-analyse er en statistisk metode, afhænger kvaliteten af analysen af, hvilken konkret metode som er brugt til at analysere data. Mange faktorer kan føre til bias, forkert fortolkning eller give ugyldige resultater. Denne afhandling præsenterer en statistisk undersøgelse af to områder indenfor meta-analyse, hvor disse problemer kan eksistere, henholdsvis: subgruppe meta-analyse og sekventiel meta-analyse.

Det første område, subgruppe meta-analyse, har kendte problemer med ecological bias. Ecological bias er en slags bias, hvor proportionen af en subgruppe i forsøget har en effekt på behandlingseffekten. Dette bias kan føre til at effekten af en subgruppe på behandlingseffekten bliver en blanding af forskellige effekter og man kan miste sin ønskede fortolkning. Dog er dette afhængig af, hvilken metode man bruger til sin analyse. Denne afhandling bidrager med en ny metode til subgruppe meta-analyse ved brug af linear mixed models. Metoden benyttes til at estimere vekselvirkningen mellem subgruppen og behandlingseffekten og kan samtidig korrigere for eventuel ecological bias. Metoden kan især benyttes til scenarier, hvor kendte metoder er utilstrækkelige.

Det andet område, sekventiel meta-analyse, ser på gyldigheden af de statistiske resultater som fås når man opdaterer sin meta-analyse. Et kendt problem er at det at opdatere en analyse vil få type-I-fejlen til at stige. Hvis man planlægger en sekventiel meta-analyse inden den påbegyndes kan denne stigning af type-I-fejlen kontrolleres. Trial Sequential Analysis er et software skrevet i Java som er designet til at håndtere gyldigheden af de statistiske resultater i en sekventiel meta-analyse. Softwaret benytter gruppe sekventielle metoder kendt fra randomiserede kliniske forsøg. En ny version af softwaret er præsenteret i denne afhandling. Den nye version indeholder flere egenskaber og har en klarer definition af prospektive og retrospektive sekventielle meta-analyser. Bidrag til softwaret er blandt andet; nye metoder til sample size beregninger, binding og non-binding futility grænser for både en- og to-sidet designs, samt flere metoder til at beregne inferens. En open-source software pakke til det statistiske software R er præsenteret i afhandlingen.

Et andet problem indenfor sekventiel meta-analyse er en specifik form for bias. Bias betinget på beslutningen om at fortsætte en meta-analyse sker når nye studier er motiveret at tidligere evidens, som kunne f.eks. være resultatet af en meta-analyse som virker lovende. Hvis de nye studier sættes sammen med de gamle i en opdateret meta-analyse vil resultatet højst sandsynligt have en

bias mod en større effekt end den sande effekt. Inspireret af metoder indenfor gruppe sekventielle forsøg, vil denne afhandling præsentere en ny estimation-smetode til at beregne punktestimatet i en opdateret meta-analyse, hvor nye studier er sat sammen med et tidligere lovende studie eller en meta-analyse. Denne estimator vil justere for dette bias.

Statement of candidate

This thesis "Statistical methods for meta-analysis" is being submitted to Macquarie University and University of Copenhagen in accordance with the Cotutelle agreement dated 21-01-2020. This work has not previously been submitted for a degree or diploma in any university.

To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself. I also certify that all information sources and literature used in this work are referenced in the thesis.

I also certify that this thesis has been written by me, and in totality, my contribution was at least 90% of the total effort required to conduct and complete this research. Specifically, the research conducted as part of this degree contributed towards three papers. Two papers were co-authored with my Australian supervisor, Professor Ian C. Marschner. The remaining paper was co-authored with my Danish supervisors, Professor Theis Lange, Professor Christian Gluud and MD Markus Harboe Olsen. In each of these cases, the contribution of the co-author was to assist with conception of the methods, and provide general supervision and feedback on the research and writing. Additionally, a software package also included in this thesis was co-authored with Markus Harboe Olsen.



25/10/2023

Anne Lyngholm Sørensen

Date

Publications and presentations

Peer-reviewed journal articles

Sørensen A. L. and I. C. Marschner (2023). Linear mixed models for investigating effect modification in subgroup meta-analysis. *Statistical Methods in Medical Research*. 2023;32(5):994-1009. doi:10.1177/09622802231163330

Submitted manuscripts

Soerensen A. L, M. H. Olsen, T. Lange and C. Gluud (2023). RTSA: An R package for the updated version of Trial Sequential Analysis. *Submitted to Journal of Statistical Software August 2023*.

Manuscripts ready for submission

Sørensen A. L. and I. C. Marschner (2023). Adjusting for conditional bias arising from a decision to continue a sequential meta-analysis. *Ready for submission to Statistics in Medicine*.

Software packages

Soerensen A. L and M. H. Olsen (2023). RTSA: “Trial Sequential Analysis” for Error Control and Inference in Sequential Meta-Analyses. R package version 0.2.1. <https://github.com/AnneLyng/RTSA>.

Conference abstracts

Soerensen, A. L. *Linear Mixed Models for interaction effects in meta-analysis*. Presented at JB Douglas Postgraduate 21st Award, University of Sydney, Sydney, December 8, 2020. Representing Macquarie University, Sydney, Australia. (Oral presentation).

Soerensen, A. L. *Sample size and trial number considerations when conducting random-effects meta-analyses*. Presented at the 63rd International Statistical Institute (ISI) World Statistics Congress, July 11-16, 2021. (Oral presentation).

Soerensen, A. L. *Prospective and retrospective sequential meta-analysis using*

Trial Sequential Analysis. Presented at the 44th Annual Conference of the International Society for Clinical Biostatistics, Milan, Italy, August 27-30, 2023. (Oral presentation).

Contents

Preface	i
Abstract	iii
Resumé	v
Statement of candidate	vii
Publications and presentations	ix
1 Introduction	1
1.1 Three problems in meta-analysis	2
Objectives	2
1.2 Thesis outline	3
2 Meta-analysis	5
2.1 Introduction	5
2.2 Meta-analysis	6
Heterogeneity in meta-analysis	8
Sample size estimation in meta-analysis	9
3 Subgroup meta-analysis	13
3.1 Introduction	13
3.2 Methods for subgroup meta-analysis	14
Confounding in subgroup meta-analysis	17
4 Trial Sequential Analysis	19
4.1 Sequential meta-analysis	19
Trial Sequential Analysis	20
4.2 Group sequential methods	21
Introduction	21
Calculating testing thresholds in sequential designs	22
4.3 Sequential meta-analysis using TSA	27
5 Computational methods	33
5.1 Type-II-error in group sequential designs	33
Futility boundaries	34
5.2 Computational methods	37
Numerical integration	37
Complete design	38

Testing	39
5.3 Final remarks on the package	39
6 Sequential conditional bias	41
6.1 Context	41
6.2 Analysis of sequential data for sequential trials	42
Conditional point estimates	43
6.3 Conditional bias due to decision making	45
Motivation and set-up for Manuscript III	47
7 Conclusions	49
7.1 Main contributions	49
7.2 Ongoing and future research	50
Subgroup meta-analysis	50
Trial Sequential Analysis	50
RTSA	51
Effect smaller than minimal clinical relevant value	51
7.3 Summary	52
8 Manuscripts	53
8.1 Manuscript I	53
Technical details I	70
8.2 Manuscript II	82
8.3 Manuscript III	113
Appendix A RTSA manual	133
Appendix B Additional results	155
B.1 Settings	155
B.2 Simulation scheme	155
B.3 Examples of simulated sequential meta-analyses	158
Prospective sequential meta-analysis	158
Retrospective sequential meta-analysis	160
B.4 Results	161
Prospective sequential meta-analysis	162
Retrospective sequential meta-analysis	163
B.5 Preliminary conclusions	164
Bibliography	167

1. Introduction

This thesis investigates three problems within the field of meta-analysis. A meta-analysis is a statistical summary of the results of multiple trials investigating very similar hypotheses. Using a weighted average of the trials' point estimates, a meta-analysis provides a summarised estimate to condense the information. Meta-analyses are described in more detail in Chapter 2.

There exist an abundance of meta-analyses published. Searching PubMed, 186,834 results of the article type *meta-analysis* are found from 1990 to September 2023. Furthermore, the number of published meta-analyses is increasing by the year as shown on Figure 1.1. These 186,834 meta-analyses might exist for different reasons. Besides the usefulness of a meta-analysis for achieving an overarching conclusion, a meta-analysis can also provide insight into whether trial results are heterogeneous, it can investigate the potential sources of the heterogeneity, it can detect subgroups that benefit most from the intervention and more.

With the large number of meta-analyses published and the different motivations, there are a number of complexities which the statistical analysis has to take into consideration. This thesis is concerned with statistical methodology for meta-analyses for two of such complexities, one concerns subgroup meta-analysis on aggregated data and the other concerns the statistical methods used for sequential meta-analysis. Within these two areas, we will look at statistical methods used for three specific problems. These three problems

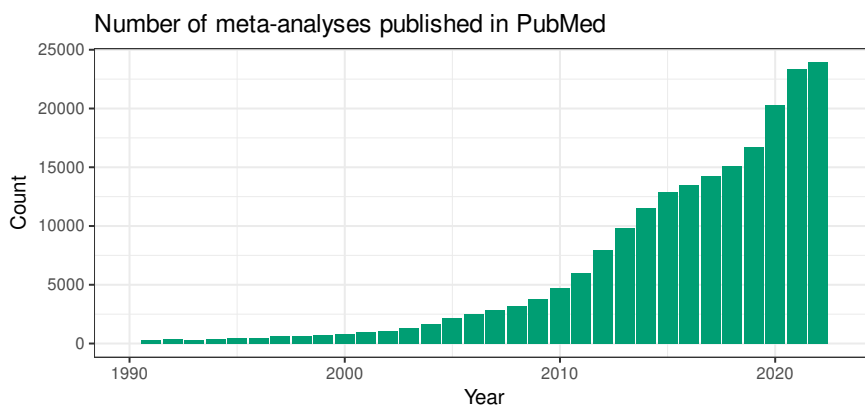


Figure 1.1: Number of publications characterised as meta-analysis per year from PubMed.

resulted in three research papers which will be presented in the latter part of the thesis and form the primary body of novel research in the thesis.

This introductory chapter will try to give a short description of the problems resulting in the thesis' objectives, whereas the remainder of the thesis will be concerned with the background and our proposed solutions to the problems.

1.1 Three problems in meta-analysis

In the first considered area of meta-analysis, subgroup meta-analysis, the first manuscript, Manuscript I, is motivated by earlier research on the topic, as presented in a paper titled “Meta-analytical methods to identify who benefits most from treatments: daft, deluded, or deft approach?” by Fisher et al. (2017). The paper considered aggregate data for subgroup meta-analysis and evaluated three common methods and their sensitivity towards ecological bias. Here the conclusion was that the interaction meta-analysis method, named the “deft” method, was the most appropriate to use for estimating effect modification in subgroup meta-analysis using aggregate data. However, the method can only accommodate data which includes all subgroups of interest. Relevant but incomplete data, which is data missing one or more subgroups of interest, cannot be part of the model. Our goal with Manuscript I was to create a model which handles ecological bias and incomplete data in a more comprehensive analytical framework.

In the second area of meta-analysis considered in this thesis, sequential meta-analysis, we had two objectives. The first objective was to investigate the type-I- and type-II-error control in meta-analyses which get updated over time. Trial Sequential Analysis (TSA) is one such method proposed by Copenhagen Trial Unit (Wetterslev et al., 2008). To be able to look at the software's ability to control false-positives and false-negatives, the method was implemented in statistical software R (R Core Team, 2020) with several updates. Thus the objective was to implement TSA in R and create an R package for others to use.

The last problem is also related to sequential meta-analysis. The motivation for the last manuscript, Manuscript III, were two published papers: “Delayed Vs Early Umbilical Cord Clamping for Preterm Infants: a Systematic Review and Meta-Analysis” by Fogarty et al. (2018) and “The Effect of Lactoferrin Supplementation on Death Or Major Morbidity in Very Low Birth-weight Infants (LIFT): a Multicentre, Double-Blind, Randomised Controlled Trial” by Tarnow-Mordi et al. (2020). Both contain a meta-analysis which include later studies that were motivated by previous promising yet insignificant meta-analyses. This creates a bias in the updated meta-analyses, when it is only updated due to its predecessor being promising. Manuscript III investigates whether we can find an estimator which adjusts for the conditional bias of continuing a promising meta-analysis.

Objectives

Based on the discussion above, we summarise our research goals into three objectives:

- Subgroup meta-analysis

1. Development of a statistical method for estimating within-population effect modification in subgroup meta-analysis. The method will use the linear mixed model setting and provides a comprehensive analytical framework for handling various data types including missing data. The method should be able to handle ecological bias.
- Sequential meta-analysis
2. Contribute to the Trial Sequential Analysis (TSA) software with new state-of-the-art methods within sequential trials and meta-analysis. Implement the revised method in statistical software R for ease of use for others and for potential simulation studies. The software implementation is then to be disseminated as a package in CRAN and a software paper along with several vignettes.
 3. Development of an analytical method to handle conditional bias in updated meta-analyses stemming from the decision to continue the analysis due to promising results. The method should aim to adjust the point estimate of the updated meta-analysis, where the bias is introduced from a decision-making process. Here it is assumed that a previous analysis (trial or meta-analysis) motivated the conduct of new trials based on a region of the test-statistic from the previous evidence. The analysis is then updated, where the original analysis is combined with the new trials. We aim to adjust the point estimate of this updated meta-analysis for conditional bias due to the decision to continue the analysis under observed promising results.

The background and proposed solutions to these objectives are presented within this thesis.

1.2 Thesis outline

This thesis is presented as a “thesis by publication” and conforms to a specific required format. The first part of the thesis presents background and discussion relevant to the interpretation of the novel research, which is presented in the latter part of the thesis in the form of three manuscripts. The three manuscripts are all in the form of journal articles that are either already published or ready for peer review.

Within this format, the thesis consists of eight chapters. The first chapter is an introduction to the remainder of the thesis, and consists of a general introduction, the objectives of the thesis, and an outline. In the style of a synopsis, this chapter is followed by a methods section. The methods section runs from Chapter 2 to Chapter 6 serving as background for the manuscripts (Manuscript I-III) appearing later in the thesis at Chapter 8 before the bibliography. The main novel contribution of this thesis is the independent research presented in these three manuscripts. A short explanation of the methods chapters is given below with comments about which chapters are of relevance to each of the manuscripts:

- Chapter 2 introduces meta-analysis in general which is relevant for all manuscripts presented in this thesis.

- Chapter 3 introduces subgroup meta-analysis, the most commonly used methods within the field, and ecological bias in aggregated data meta-analysis. This chapter is an introduction to Manuscript I.
- Chapter 4 presents Trial Sequential Analysis, a method for sequential meta-analyses. Sequential meta-analyses are meta-analyses that have been or will be updated in their lifetime. As the method is an adaptation of group sequential designs for single trials, most of the chapter is concerned with theory for group sequential designs. This chapter is part of the introduction to both Manuscript II and Manuscript III.
- Chapter 5 describes power calculations in group sequential designs and a selection of the computational methods used in the creation of the RTSA package. This chapter serves as the final chapter for introducing Manuscript II. Supplemental material to this manuscript is found in the appendices.
- Chapter 6 introduces inference in sequential designs given the background theory in Chapter 4. The chapter concludes with highlighting the problems with conditional bias due to the decision to continue a sequential meta-analysis. This chapter serves as the last chapter introducing Manuscript III and finalises the methods chapters.

The second last chapter of the thesis, Chapter 7, contains a general conclusion, the main contributions and planned future research. The last part of the thesis is referred to as Chapter 8 but is the majority of the material in the thesis, incorporating the three manuscripts in journal format. Appendix A contains the manual for the RTSA package and Appendix B contains additional results not provided by the manuscripts relating to Manuscript II.

2. Meta-analysis

All of the manuscripts underpinning this thesis are within the field of meta-analysis. This chapter will provide an introduction and overview of the basic theory and notation used subsequently in the thesis. We start with a brief introduction to meta-analysis and a review of some topics that will be important for the research presented later in the thesis, including heterogeneity assessment and sample size determination.

The material covered in this chapter should provide the reader with a basic understanding of meta-analysis. For more comprehensive background literature on meta-analysis, we recommend the books “Introduction to Meta-Analysis” by Borenstein et al. (2009) and “Methods for Meta-Analysis in Medical Research” by Sutton et al. (2000).

2.1 Introduction

In many fields of research, different trials are testing the same hypothesis. Meta-analysis is used to obtain a summary of the results of these different trials often expressed by a point estimate, confidence interval and p-value. It synthesises the results of the trials to provide an estimate representing an overall direction and size of the effect of interest. In this way the results of multiple studies can be synthesised and the evidence is consolidated by reducing the number of conclusions from many to one. The result of a meta-analysis can suggest if more information is needed. Hence the result of the meta-analysis may not only be a conclusion provided the current available trials, but it may also be used an argument for more trials to be planned or where it seems to be most relevant to gain more knowledge. Another reason for conducting a meta-analysis is the expected gain in power. Rare event trials or trials investigating small participant groups may not always be able to recruit as many participants as required for achieving a desired level of power of their statistical analysis. Combining trials in a meta-analysis often results in a higher level of power compared to the trials individually. However, this depends on the level of heterogeneity of the trial results. Heterogeneity describes the between-trial variation of the trial results used in a meta-analysis. Heterogeneity will be described in more detail later in this chapter.

Data used for meta-analysis comes in two forms; individual participant data or aggregate data. Individual participant data is the raw data on each participant individually. Aggregate data provides data at trial level such as point estimates, standard errors and maybe some descriptive data of baseline characteristics or such. We will in this thesis only consider meta-analysis on aggregate data.

2.2 Meta-analysis

An aggregate data meta-analysis point estimate is calculated as a weighted average of the included trials' point estimates. Suppose we have K trials available in the meta-analysis. Let $k = 1, \dots, K$ be the identifier of trial k . Each trial reports an estimate of the intervention effect y_k and an estimate of the associated variance (squared standard error) s_k^2 . Here y_k can be a mean difference, a log odds ratio or another common treatment effect measure. The inverse of the variance is called information - the smaller the trial's treatment effect standard error, the higher the information. The information is often used as weights in the weighted average, as it is the weighting which minimises the variance of the pooled average. However, there are a library of other methods used to calculate the weights. These include the Mantel-Haenszel method or the Peto odds ratio method for binary outcome data (Borenstein et al., 2009). To calculate the inverse-variance weights, the estimate of the variance s_k^2 is used as a plug-in estimate for the actual variance σ_k^2 . Denoting the weights as w_k , the formulae for the meta-analysed point estimate, variance and inverse-variance weights are then defined as:

$$\hat{\theta} = \sum_k \frac{y_k \cdot w_k}{\sum_k w_k}, \quad \text{and,} \quad \text{var}(\hat{\theta}) = \frac{1}{\sum_k w_k} \quad \text{with} \quad w_k = \frac{1}{s_k^2}. \quad (2.1)$$

This is known as a fixed-effect meta-analysis. A fixed-effect meta-analysis assumes all trials are estimating the same true intervention effect θ and that the study-specific treatment effect estimates have a normal distribution (either exactly or asymptotically), thus $y_k \sim \mathcal{N}(\theta, \sigma_k^2)$. With the assumption of normality, one can calculate confidence intervals, test-statistics, and p-values using the normal distribution.

A meta-analysis can be visualised by a forest plot as shown on Figure 2.1. Here the results of the trials are plotted as points with the variability shown by 95% confidence intervals. The size of the points are relative to the weight of the trial result. In this example the weights are homogeneous due to the trials simulated for this example being of equal size with no heterogeneity in the simulated data. The pooled effect is shown in the bottom of the plot as a diamond with the center of the diamond being the estimated intervention effect and the horizontal ends its variability.

The trials included in a meta-analysis will often come from different populations or have other underlying differences. This can cause heterogeneity in the study-specific results. Inclusion and exclusion criteria, for the trials to be included in the meta-analysis, are used to minimise the heterogeneity. However, it is often still expected and a random-effects model can be used instead of the fixed-effect model to model this heterogeneity. The random-effects model is fitted similarly to the fixed-effect model expect for the inclusion of a term for the between-study variance, τ^2 , which is used in the weights. This changes the formulae from equation (2.1) to:

$$\hat{\theta}^R = \sum_k \frac{y_k \cdot w_k^R}{\sum_k w_k^R} \quad \text{and} \quad \text{var}(\hat{\theta}^R) = \frac{1}{\sum_k w_k^R} \quad \text{with} \quad w_k^R = \frac{1}{\tau^2 + s_k^2}. \quad (2.2)$$

Including τ^2 in the weights increases the variance estimate of the pooled intervention effect compared to the fixed-effect meta-analysis. It will shift some of

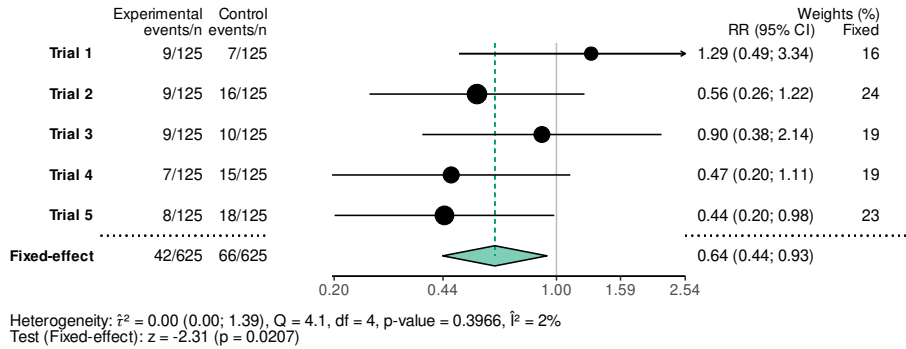


Figure 2.1: Fixed-effect meta-analysis forest plot based on simulated data without heterogeneity.

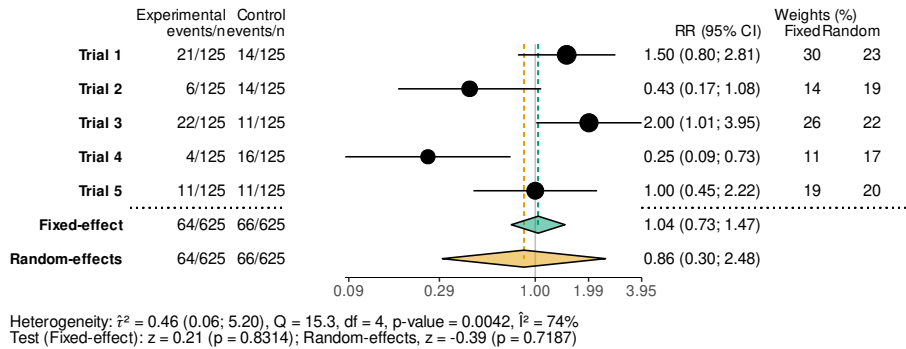


Figure 2.2: Random-effects meta-analysis forest plot based on simulated data with heterogeneity.

the weight from low variance trials to trials with a higher variance. This results in the random-effects pooled intervention effect being influenced more by all trial results compared to the fixed-effect pooled intervention effect. Figure 2.2 visualises a random-effects meta-analysis on a forest plot. The weights of both the fixed-effect and the random-effects model are presented on the figure showing that some of the weight from the trials with the most information shifts to more variable trials in the random-effects meta-analysis compared to the fixed-effect meta-analysis. The increased variability is also shown by comparing the two diamonds, where the random-effects intervention effect diamond is wider than the fixed-effect diamond.

The random-effects meta-analysis also differs from the fixed-effect meta-analysis in assumptions and thus interpretation. The trials are no longer assumed to estimate the same intervention effect. Instead it is assumed that all trials have their own true intervention effect θ_k which comes from a distribution centred around a true average effect θ . Thus, the study-specific treatment effects, $y_k \sim \mathcal{N}(\theta_k, \sigma_k^2)$ where $\theta_k \sim \mathcal{N}(\theta, \tau^2)$ with τ^2 being the measure of the between-trial variation.

Whether a fixed-effect or random-effects meta-analysis model is used can de-

pend on several factors. In principle one can pre-specify the use of a fixed-effect meta-analysis. In this scenario one assumes that all studies used will be almost identical and that we want a common effect size representing just the populations represented in the trials that we will use for the meta-analysis. Others might prefer to be more data-driven in their decision of choosing between fixed-effect or random-effects models by testing whether or not heterogeneity exists. This method is, however, problematic as we can have a poor precision of the τ^2 estimate. It is known that τ^2 is underestimated when only having a handful of trials included in the meta-analysis. Hence the strategy is at most useful for concluding the presence of heterogeneity rather than concluding no presence. Figure 2.2 compares a fixed-effect meta-analysis to a random-effects meta-analysis. Both estimated pooled effects can be valid based on the desired interpretation of the analysis. In this thesis we will not study the choice between fixed effects and random effects meta-analysis in detail, but rather we will consider new meta-analysis methodology within the context of these two approaches. The next section introduces different methods for estimating heterogeneity in meta-analysis.

Heterogeneity in meta-analysis

Heterogeneity can be quantified in a meta-analysis. The usual measure of heterogeneity is $\hat{\tau}^2$ which is calculated using the Q statistic defined as $Q = \sum_k w_k \cdot (\hat{\theta}_k - \hat{\theta})^2$ (Borenstein et al., 2009). This statistic is also known as Cochran's heterogeneity and quantifies the variability of the point estimates of the trials relative to the variability of the trial results. A method to estimate the heterogeneity τ^2 is based on the Q statistic and defined by:

$$\hat{\tau}^2 = \max\left(0, \frac{Q - (k - 1)}{\sum w_k - \sum w_k^2 / \sum w_k}\right).$$

The formula is a moment-estimator of τ^2 and is often chosen for its simplicity. It is based on the first moment of the Q statistic, $E[Q]$, which depends on τ^2 . As the formula was derived by DerSimonian and Laird (DerSimonian et al., 1986), it is often called the DerSimonian-Laird estimate of heterogeneity. One can plug-in the estimate of τ^2 to calculate the weights in (2.2) used in estimating the random-effects pooled estimate and associated variance. There exist other methods for estimating τ^2 using e.g. maximum likelihood estimation or restricted maximum likelihood estimation (Seide et al., 2019).

To test for heterogeneity of the trials' estimated effects, one can use that the Q statistic under the null hypothesis of no heterogeneity will follow a χ^2 distribution with degrees of freedom equal to $K - 1$ (Borenstein et al., 2009). However the estimation of the between-trial variance and hence the Q statistic is unstable for a small number of trials. For this reason there are multiple other estimators for the variance of the pooled estimate been proposed. One example is the Hartung-Knapp-Sidik-Jonkman (HKSJ) estimator defined as (IntHout et al., 2014):

$$\text{var}(\hat{\theta}^R) = \frac{w_k^R \cdot (\hat{\theta}_k - \hat{\theta}^R)^2}{(k - 1) \sum_k w_k^R}.$$

This estimator of the variance of the pooled effect uses the same expression for τ^2 as before, hence the DerSimonian-Laird estimator. The estimate of the

variance is based on a moment-estimator for the t -distribution instead of the normal distribution. Hence when using the HKSJ adjustment, one assumes the t distribution, $\theta_k \sim t(\theta, \tau^2, \nu)$, where ν is the degrees of freedom equal to the number of trials K minus 1. Methods to derive confidence intervals for the estimate of τ^2 have been created to express its uncertainty. Two of these methods, Q-profile method and the Biggerstaff-Tweedle method (Viechtbauer, 2006; Biggerstaff et al., 1997), are used in the software package later presented in this thesis.

To express heterogeneity on different scales, other metrics have been created. Inconsistency I^2 by Higgins et al. (2002) and diversity D^2 by Wetterslev et al. (2017) are two measures for quantifying the impact of between-trial variation on the total variance estimate of the pooled effect size. The measures are both expressing the size of heterogeneity relative to the total size of variation as a percentage. The total variation is a sum of the heterogeneity τ^2 and the within-study variability (sampling error). The two measures I^2 and D^2 are different as they have different estimates for the within-study variability. Defining first inconsistency I^2 , a moment-based sampling error estimator is used for the within-study variation, here denoted σ_M^2 :

$$I^2 = \frac{Q - (k - 1)}{Q} = \frac{\tau^2}{\tau^2 + \sigma_M^2} \quad \text{where} \quad \sigma_M^2 = \frac{\sum w_k(k - 1)}{(\sum w_i)^2 - \sum w_i^2}. \quad (2.3)$$

Here σ_M^2 is called a ‘‘typical’’ within-study variance estimate. In comparison diversity D^2 calculates the proportion between the between-trial variance and the total variance as:

$$D^2 = \frac{\tau^2}{\tau^2 + \sigma_D^2} \quad \text{where} \quad \sigma_D^2 = \frac{\tau^2 \cdot \text{var}(\hat{\theta}^R)}{\text{var}(\hat{\theta}^R) - \text{var}(\hat{\theta})}. \quad (2.4)$$

Hence D^2 is the exact estimated proportion between the between-trial variation and the sum of variances (between and within trials).

Sample size estimation in meta-analysis

The research presented subsequently in this thesis contains theory involving the needed number of participants to achieve a specific power in a given meta-analysis. Power is the probability of rejecting the null hypothesis under the assumption that the null hypothesis is false given a certain level of the intervention effect. It depends on the size of the anticipated intervention effect, its variance, the number of participants in the meta-analysis, and the testing strategy. The desired level of power in a trial is therefore used as a design parameter to calculate the required number of participants. Many meta-analyses will not have an a priori or posteriori sample size calculation, but when using sequential methods, as is central to the methodology presented here, it will be necessary. This section will review sample size estimation when the testing strategy is to test the null hypothesis once. Chapter 5 describes sample size estimation for sequential meta-analysis where the null hypothesis is tested more than once.

Sample size estimation can be carried out prior to the meta-analysis to design meta-analyses with a specific level of power. Both power and sample sizes are also sometimes calculated retrospectively to calculate the achieved

power and, if necessary, the additional required participants to achieve the desired level of power of an existing meta-analysis. Here we will focus on methods for calculating the number of participants in a yet-to-be-conducted meta-analysis.

We use the term required information size (RIS) for the required sample size in a meta-analysis (Pogue et al., 1997; Wetterslev et al., 2008). The sample size calculation can be performed using the usual formula for normally distributed effects if one wishes to use a two-sided test with the fixed-effect meta-analysis model (Pogue et al., 1997):

$$RIS = 4 \cdot (z_{1-\alpha/2} + z_{1-\beta})^2 \cdot \frac{\nu}{\theta^2}. \quad (2.5)$$

Here, RIS, required information size (sample size) can be split across the desired number of trials included in the meta-analysis to achieve the desired level of power. In the formula, z_x is the x^{th} quantile of the standard normal distribution where α is the type-I-error and $1 - \beta$ is the power. The formula can be used for both continuous and binary outcomes. For continuous data, ν is the expected variance and θ the expected mean average. For binary data will $\nu = p_A \cdot (1 - p_A)$ and $\theta = p_I - p_C$, where $p_A = (p_I + p_C)/2$ is an average of the expected probabilities of event in the intervention group, denoted by p_I , and the control group, denoted by p_C .

To account for heterogeneity in the sample size calculation Wetterslev et al. (2017) presents two formulas based on the relative measures of heterogeneity in the sample, I^2 and D^2 :

$$RIS_{I^2} = \frac{1}{1 - I^2} \cdot 4 \cdot (z_{1-\alpha/2} + z_\beta)^2 \cdot \frac{\nu}{\theta^2}, \quad \text{and}$$

$$RIS_{D^2} = \frac{1}{1 - D^2} \cdot 4 \cdot (z_{1-\alpha/2} + z_\beta)^2 \cdot \frac{\nu}{\theta^2}.$$

Both methods will increase the sample size relative to the size of I^2 and D^2 . As $D^2 \geq I^2$, the sample size based on D^2 will be equal to or larger than the one based on I^2 . The method, Trial Sequential Analysis, uses the diversity-adjusted required information size, which is abbreviated to DARIS (Thorlund et al., 2011).

It was shown by Kulinskaya et al. (2013) that to achieve a given level of power, it is not always sufficient to only let the sample size increase - a minimum number of additional trials may also be required. Kulinskaya et al. (2013) provides formulae for calculating the remaining number of participants and the minimum number of additional trials in an already existing meta-analysis to achieve a specific power. These formulae can be interpolated to calculate the required number of participants and trials before observing any data. Let θ be the intervention effect of interest, which has estimator $\hat{\theta}$ with $\text{var}(\hat{\theta})$ its expected variance, τ^2 the expected level of heterogeneity, α and β be respectively the type-I and type-II-error rate and z_x be the quantile from the standard normal distribution at x . When using the Wald test, we have that:

$$\frac{\theta}{\sqrt{\text{var}(\hat{\theta})}} = z_{1-\alpha/2} + z_{1-\beta}, \quad \text{where} \quad \text{var}(\hat{\theta}) = \left(\sum_k \frac{1}{2 \cdot \sigma_k^2/n_k + \tau^2} \right)^{-1}.$$

As the expression of the right is bounded by the number of trials K , we will be able to achieve the desired power, $1 - \beta$, when:

$$\tau^2 < \frac{\theta \cdot K}{(z_{1-\alpha/2} + z_{1-\beta})^2}.$$

To achieve this we need to find a K for which the above inequality holds. Given the equation a minimum number of trials K is needed but there is no unique solution for K . In theory, one will have a well-powered meta-analysis by letting $K \rightarrow \infty$ which then affects the number of participants per trial n_k , where $n_k \rightarrow 2$. Finding the optimal K depends on whether one prioritises the individual trial power or wishes to minimise the total sample size of all the trials combined. The required number of participants per trial is calculated by:

$$n_k = \frac{4 \cdot \sigma^2}{\frac{\theta \cdot K}{(z_{1-\alpha/2} + z_{1-\beta})^2} - \tau^2}.$$

Both the original formulas from Kulinskaya et al. (2013) and the ones just presented are implemented in computational methodology presented later in the thesis.

3. Subgroup meta-analysis

This chapter presents the rationale behind subgroup meta-analysis together with some of the field’s most common methods. Each of these methods have specific strengths and drawbacks, especially with respect to an important type of confounding that can occur in subgroup meta-analysis, called ecological bias. This type of bias is also known as aggregation bias (Riley et al., 2020). The concept of ecological bias will be presented here along with important concepts in subgroup meta-analysis. This chapter serves as an introduction to the first piece of original research which is presented here as published in the paper: “Linear Mixed Models for investigating effect modification in subgroup meta-analysis” (Soerensen et al., 2023a).

3.1 Introduction

The effect of a treatment on a given outcome might vary between patient groups. Some patients might experience a larger effect of an intervention, whereas others might experience a smaller or no effect. Identifying differential effect of treatment by group can provide patients with personalised treatment. Thus, patients can receive better care by targeting the treatment to the group of patients for whom the treatment is considered beneficial. A subgroup is the name applied to a group of participants possessing a common profile. The study sample is split into a collection of subgroups based on a characteristic of the participants of the trial which either naturally, or by design, separates the sample into nominal categories. These categories can be based on any baseline patient characteristic, such as e.g. smoking status, sex, or age-group. If there is differential intervention effects based on subgroup, the characteristic, e.g. smoking status, defining the subgroup is called an effect modifier. From a purely statistical perspective it is an interaction effect of the subgroup covariate on the effect of intervention. Effect modification is visualised in Figure 3.1. Figure 3.1 Panel 1) shows the relationship between the treatment A , outcome Y and how this association is modified by subgroup G . The figure shows that the allocation of treatment A is not dependent on subgroup G but that the effect stemming from treatment A on outcome Y is modified by subgroup G .

Many randomised controlled trials (RCTs) will not have enough power to investigate whether there is differential intervention effect based on subgroup. Often, the main aim of an RCT is to show effect of intervention compared to control on a more general population. For this reason subgroup effects are likely to be investigated using meta-analysis, where one can combine multiple studies to achieve an increase in power compared to analysing the subgroup effects in a single study. A branch of meta-analysis is dedicated to methods

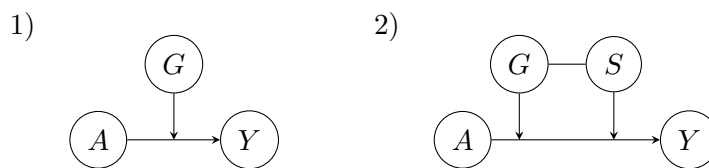


Figure 3.1: Effect modification by subgroup G on the effect from treatment A on outcome Y . Plot 1): Effect modification. Plot 2): Effect modification with study-level confounding. S denotes study.

for investigating the effect of subgroup on the intervention effect and is called subgroup meta-analysis. Besides identifying differences in intervention effect, subgroup meta-analysis can be used to investigate sources of heterogeneity in the overall meta-analysis, which is a meta-analysis for the entire participant group. Subgroup meta-analysis can aid in detecting whether heterogeneity stems from specific participant groups.

To investigate presence of interaction effects on the intervention effect on outcome usually one of three types of subgroup meta-analysis methods is used. These are: interaction meta-analysis, subgroup specific meta-analysis, and meta-regression. We have in Manuscript I proposed an additional method for subgroup meta-analysis to handle shortcomings of the three methods. The three usual methods are presented below, whereas our proposed model is presented in Manuscript I.

3.2 Methods for subgroup meta-analysis

Let $y_{j,k}$ be the estimated treatment effect in study k subgroup j , where $k = 1, \dots, K$ and $j = 1, \dots, J$. All methods presented in this chapter and in Manuscript I use aggregate data, which is why we define $y_{j,k}$ on study level and not on individual patient level. We are interested in the estimate of $\Delta_{j,j'}$ which is the difference in treatment effect between subgroup j and subgroup j' , where $j \neq j'$. We will consider the scenario where we only have two subgroups for simplicity, thus we denote $j = \{1, 2\}$ and $\Delta_{2,1}$ as Δ , the estimate of interest. Besides a brief theoretical introduction of the methods, we will present the three existing methods on toy-data graphically. In our toy-data, four studies are simulated. Three studies investigate the effect of treatment on outcome for two subgroups, the fourth study only investigates the effect of treatment on outcome for one subgroup. The true effect modification was set to $\Delta = -3$ in the simulated data. All studies and the subgroup specific treatment effects are visualised using a forest plot in Figure 3.3. An introduction to forest plots was given in Chapter 2.

The first method, interaction meta-analysis, estimates Δ in a two-step procedure (Fisher et al., 2017; Fisher et al., 2011). First the difference in treatment effect between the subgroups is calculated per study. We denote the difference δ_k . Then, these study specific estimates are pooled in a meta-analysis as written in (3.1). Here w_k denotes the weight of study k where a common weighting method is the inverse of the variance, in which case w_k is reciprocal of the

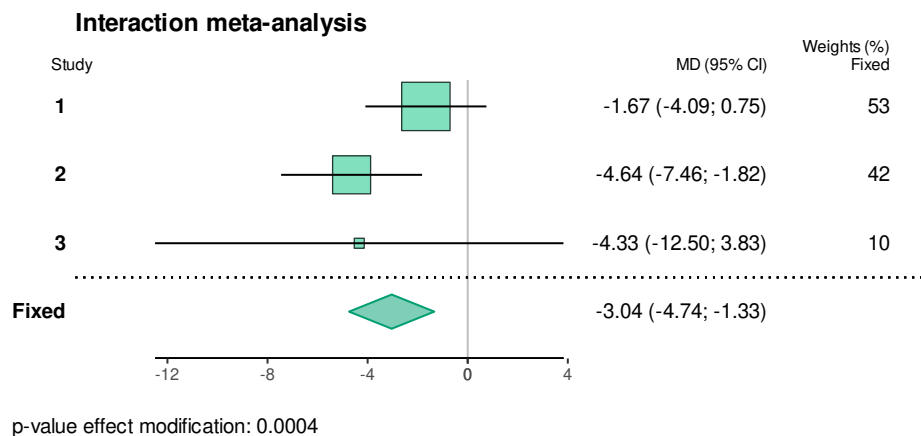


Figure 3.2: Interaction meta-analysis on three out of four toy-example studies. The fourth study could not be included as it only included participants from subgroup 1. MD stands for mean difference. A fixed-effect model is used for the meta-analysis using inverse-variance weighting. The data is simulated with no heterogeneity.

squared standard error:

$$\begin{aligned}\delta_k &= y_{1,k} - y_{0,k}, \\ \hat{\Delta} &= \sum_k \frac{\delta_k \cdot w_k}{\sum_k w_k}.\end{aligned}\quad (3.1)$$

The method is visualised in Figure 3.2, where a fixed-effect meta-analysis is used. The interaction meta-analysis will produce an unbiased estimate of the within-patient effect modification, which is often the metric we are interested in. We will discuss the concept of within-patient effect modification in the next Subsection. A drawback of the method is the incapability to include studies that do not have data on both subgroups of interest, which can cause the representation in the subgroup meta-analysis to be limited. In terms of heterogeneity, the interaction meta-analysis method can investigate whether there is heterogeneity between the interaction effects.

The second method, the subgroup specific meta-analysis method, is also a two-step procedure. In the first step, the two subgroups' specific treatment effects are pooled per subgroup which we denote below by δ_j in (3.2). Thus two initial meta-analyses are created. In (3.2), $w_{j,k}$ is the weight specific to the estimate of the subgroup specific treatment effect in subgroup j study k . The second step calculates the difference of the two pooled estimates:

$$\begin{aligned}\delta_j &= \sum_k \frac{y_{j,k} \cdot w_{j,k}}{\sum_k w_{j,k}}, \\ \hat{\Delta} &= \delta_1 - \delta_0.\end{aligned}\quad (3.2)$$

The method is visualised on a forest plot in Figure 3.3. Here fixed-effect models are used to calculate two meta-analyses, one per subgroup. The effect modifica-

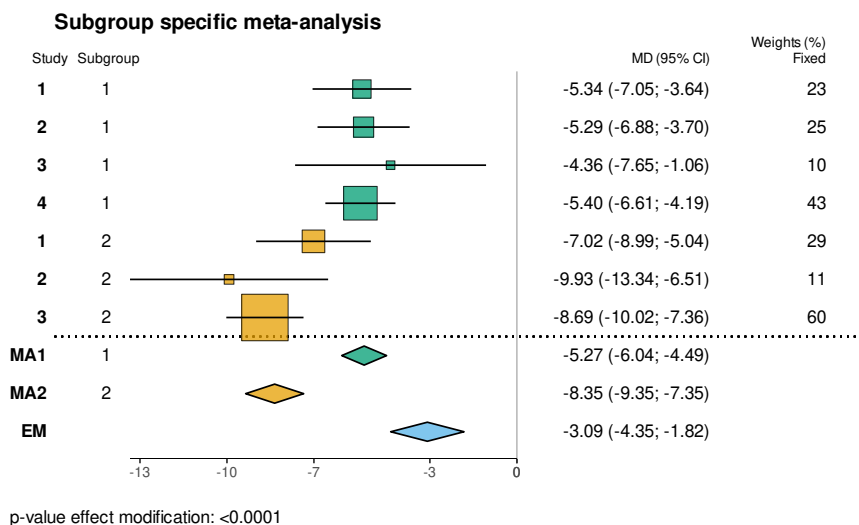


Figure 3.3: Visualisation of subgroup specific meta-analysis. The weight column on the right contains both of the two meta-analyses, which are plotted here on the same forest plot. MA1 stands for meta-analysis for subgroup 1, MA2 for meta-analysis for subgroup 2 and EM for effect modification. The sizes of the boxes are relative to the weight of the trial in the meta-analysis. All studies are included in this method.

tion is then calculated as the difference. We find that the effect of the subgroup on the treatment effect on outcome is similar to the estimated effect using interaction meta-analysis. A benefit of the subgroup-specific meta-analysis is that it will include all studies. However, it will be affected by what is called ecological bias, which is presented in the next Subsection. The subgroup-specific meta-analysis can investigate the heterogeneity within each subgroup, as the first step of the method is to fit two separate meta-analysis. Thus it can aid in detecting whether heterogeneity is more present in one subgroup compared to the other.

The last common method, meta-regression, models the study specific treatment effect y_k via a weighted regression of y_k with the ecological level as an independent variable. The ecological level is defined as the proportion of subgroup level $j = 1$, $p_{j=1,k}$ as done in Fisher et al. (2017), leading to the model:

$$y_k = b + \hat{\Delta} \cdot p_{j=1,k} + \varepsilon_k.$$

Here the effect modification is estimated as a slope parameter $\hat{\Delta}$ in the regression model, where b is the intercept and ε_k is the noise where $\varepsilon_k \sim \mathcal{N}(0, 1)$. The regression is usually weighted by the information, which is the inverse-variance weights (see Chapter 2). The method is visualised in Figure 3.4. As the method is not sensitive to studies where only one of the subgroups are present, all studies can be used for meta-regression. Heterogeneity in this scenario is estimated as the variance between the study intervention effects on the entire study population. Thus it can not be used for investigating sources of heterogeneity.

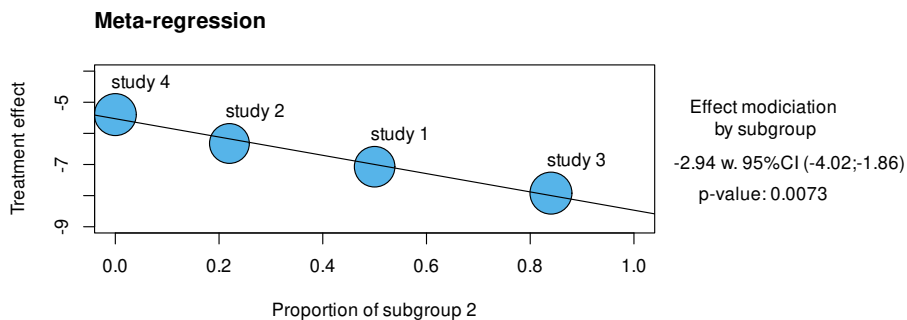


Figure 3.4: Visualisation of meta-regression. The estimated effect modification is printed on the right hand side of the plot with a 95% confidence interval and p-value. The sizes of the points correspond to the weights which are almost identical in this toy-example.

In the toy-example provided, the three methods gave almost similar estimates of the effect modification of subgroup. This is not always the case as their sensitivity to ecological bias differs. Ecological bias, which is a key concept in subgroup meta-analysis, is discussed in the next subsection.

Confounding in subgroup meta-analysis

When estimating effect modification by subgroup using the above methods, the type of association estimated can differ depending on which method is chosen. This can happen, as the three methods can, in theory, estimate different types of associations between the subgroup and intervention effect. This is not based on the data used, but rather on how the methods are constructed. For example when using meta-regression, the estimated intervention effect will reflect the association between the ecological-level of the subgroup, which is the proportion of the subgroup, on the treatment effect on the outcome. In comparison the interaction meta-analysis method will estimate the individual-level association. The two types of association, ecological-level and individual-level, might be the same but this is not always the case.

The two types of association can be clearly expressed via an example. Suppose that we are interested in the effect of gender (the subgroup) on the effect of having work experience for more than 10 years (the treatment) on salary (the outcome). A source of the data could stem from different work-places (the studies) to estimate the potential effect modification of gender. For this research question, we might use any of the three methods just presented. Based on the different work places used, we will most likely have different prevalence of a given gender, as certain fields of labour can be dominated by a specific gender. An ecological-level association informs whether the size of the prevalence of a gender across workplaces will influence the change in salary. An individual-level effect modification informs whether there is a difference on the effect of work experience on salary based on the individual's gender, thus within a workplace whether the gender will influence the change in salary. As the ecological-level

of the subgroup is directly related to the specific study, the association is also known as the study-level association. Figure 3.1 Panel 2) visualises this scenario where the prevalence of subgroup can be different in the included studies and the prevalence affects the outcome. Hence if the allocation of subgroup G (gender) is dependent on study S (workplace) and S (workplace) affects Y (the change in salary - path A to Y), then we have ecological bias.

The meta-regression method uses the ecological level of the subgroup in its model, which is why it is estimating an ecological-level association. In contrast the interaction meta-analysis will target the individual-level association. This is because, while the data used in the interaction meta-analysis may be aggregated measures and not on individual scale, the effect of the ecological-level on the subgroup effect is removed by using the differences in treatment effects by subgroup within each study. The within-study differences are not affected by the ecological level of the subgroup, as one compares the subgroups under equal ecological levels. The third method, subgroup-specific meta-analysis, will estimate a combination of the two associations.

It is often of interest to estimate the individual-level association in subgroup meta-analysis. If the ecological-level association is fully explained by the individual-level association then the three methods will provide the same estimate. However, this is not always the case. Maybe for this reason is the difference between ecological-level association and individual-level association often called ecological *bias*. We are in Manuscript I interested in the individual-level association and will consider ecological effects as a bias which we name study-level confounding.

Provided with the concept of ecological bias, we can yet again compare the three methods. As mentioned, the set of included studies can differ for the three methods. Interaction meta-analysis can only include studies that include all subgroups of interest. Meta-regression and subgroup-specific meta-analysis does not have this requirement. For this reason, they might seem to be the obvious choice of method. However, both the subgroup specific meta-analysis and meta-regression are affected by the ecological level of the subgroup (Fisher et al., 2017; Fisher et al., 2011). This is not a problem if the purpose of the analysis is to have the interpretation of their estimate on a ecological level. However, as said before it is more common to be interested in what is called individual-level effect modification, and using meta-regression or subgroup specific meta-analysis can be problematic for this purpose. In Fisher et al. (2017), the recommendation is therefore only to use interaction meta-analysis.

In Manuscript I, we propose a new method that builds on the idea of interaction meta-analysis but can include all studies and adjust for ecological bias. The proposed method is based on linear mixed models. Using a linear mixed model allows for modelling of both study specific random-effects and subgroup specific random-effects which can be beneficial for investigating the sources of heterogeneity in the study results. It further allows to adjust for ecological bias. Linear mixed models have been used in other contexts of meta-analysis for modelling individual participant data or modelling the heterogeneity in aggregate data. While using a linear mixed model is rather new in subgroup meta-analysis for aggregate data and offers a greater flexibility in terms of modelling, the key contribution of the manuscript is its handling of ecological bias and missing data. This chapter finalises the introduction to Manuscript I.

4. Trial Sequential Analysis

This chapter introduces Trial Sequential Analysis (TSA) which is a method intended for type-I- and type-II-error control in a sequential meta-analysis. Both Manuscript II and III are concerned with sequential meta-analyses and some of the background theory for both manuscripts will be presented in this chapter. A brief introduction to sequential meta-analysis will be presented first before TSA is introduced. As TSA is an adaptation of group sequential methods for clinical trials to meta-analysis, this chapter will include, and primary focus on, the theory of group sequential methods for trials. While TSA is mostly relevant for Manuscript II, some of the background theory for group sequential trials is also used in Manuscript III. The chapter concludes with a section on how group sequential methods for single trials and sequential meta-analysis can be tied together and for which meta-analyses TSA may be used.

4.1 Sequential meta-analysis

We define a sequential meta-analysis to be a meta-analysis which is planned to or has been updated over time. A common reason for updating a meta-analysis is the publication of new trials happening after the preceding meta-analyses. Updating the meta-analysis will provide up-to-date evidence and has been recommended (Elliott et al., 2017). The sample size of the meta-analyses will in most cases increase as new information is added. The opposite is also possible, trials included in earlier meta-analyses may be removed from the updated ones due to e.g. newly found risk of bias. This can cause the sample size in the updated meta-analysis to decrease. In this thesis, we will primarily consider sequential meta-analyses where all previous trials are included in the updated meta-analysis. Thus scenarios where the sample size increases. Figure 4.1 depicts this scenario from both a sequential meta-analysis and sequential trial point-of-view. The sequential meta-analysis perspective is described in the parentheses. We will return to the figure in Section 4.2.

Updating the meta-analysis often includes repeating the null hypothesis test on the updated data. This is known to increase the type-I-error (Armitage et al., 1969), hence the p-value will be invalid and so will the accompanying confidence interval. TSA is a method which tries to solve type-I- and type-II-error problems in sequential meta-analyses by using group sequential methods created for clinical trials. Other methods have been proposed to solve the problems of invalid p-values and confidence intervals. Living Systematic Review (LSR) is one such method or methodology, which removes decision making from the sequential meta-analysis. This means that while the naively calculated confidence intervals and p-values can be calculated as a part of a

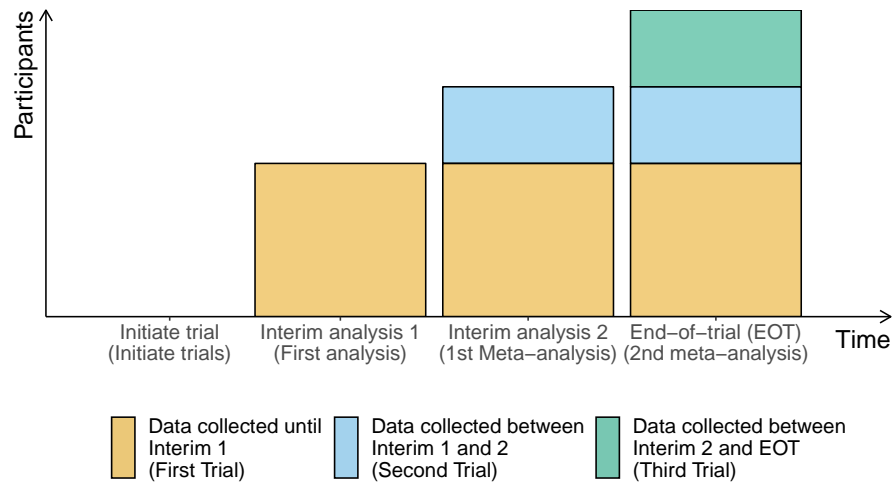


Figure 4.1: A visualization of the accumulating data in a group sequential trial. Parentheses illustrate the meta-analysis perspective.

LSR, they are not meant to be used for any decision making (Simmonds et al., 2017; Elliott et al., 2017). E-values and e-variable-based confidence intervals is yet another method. Here e-values are used, abandoning the traditional p-values and a game-theoretic interpretation is used instead of the frequentistic (ter Schure et al., 2019). We wish to stick to the frequentistic methodology as we want to investigate designs and to analyse type-I- and type-II-error rates. TSA uses frequentistic methods where we will be able to evaluate these types of errors.

It is worth noticing that there exist more methods than TSA aimed for modelling sequential meta-analyses. Frequentistic sequential meta-analysis has also been presented in Whitehead (2002) and Higgins et al. (2010). As TSA, these methods use stopping boundaries inspired by group sequential methods for trials. In the latter paper (Higgins et al., 2010), they further introduce how to add uncertainty about the heterogeneity parameter τ^2 in their sequential procedure by including a prior distribution on the heterogeneity. Formal comparison of TSA to other methods is part of future research. The research goal of Manuscript II is to be able to implement TSA in R.

Trial Sequential Analysis

TSA was created by Copenhagen Trial Unit as a software application written in Java (Thorlund et al., 2011). The method tries to correct for the increase in type-I-error by only considering a hypothesis test to be significant when the test statistic crosses an adjusted threshold. Most often will the adjusted testing thresholds be more conservative than the naive. The adjusted testing thresholds are also known as stopping boundaries. The calculation of these boundaries is similar to how stopping boundaries are calculated in group sequential trials. The following sections will present the theory by which TSA calculates its stopping boundaries.

4.2 Group sequential methods

The theory of group sequential trials is rather rich. To decrease the number of pages of this PhD thesis, we will focus on the theory on which Manuscript II and III depend the most. For Manuscript II, which is a software paper, everything presented below is of relevance as all of it is implemented in the software. For Manuscript III, it is primary the understanding of the joint distribution of the sequence of test statistics which is of importance.

Introduction

Group sequential methods are originally intended for single experiments where the conductor wishes to be able to stop the experiment early before reaching the final sample size. Reasons for stopping early are e.g. if there are early significant findings of superiority of the intervention compared to control. This is achieved by testing the hypothesis of interest multiple times as the data accumulates, but in a controlled manner that does not increase the risk of getting a false-positive (Wald, 1947; Barnard, 1946). To control the type-I-error, the null hypothesis is tested against adjusted testing thresholds depending on the significance level of the trial. Suppose we have a trial that wants to test their hypothesis $\mathcal{H}_0 : \theta = 0$ three times; Twice before having collected of all their data, and one last time when all data is collected if they fail to reject the hypothesis at any of the earlier analyses. Here θ could be a mean difference or the logarithm of an odds ratio. The analyses during the trial is called interim analyses and the time of the final analysis is called end-of-trial. End-of-trial will be when the sample size is reached for which the trial will have a set level of power. The planned sample size required in a group sequential trial is larger than the standard trial with no interim analyses which we shall see in Chapter 5.

For the three tests planned, we consider a set of the test statistics $\mathbf{Z}_3 = (Z_1, Z_2, Z_3)$ if the trial runs all the way to end-of-trial. For a normally distributed outcome the test statistic could be $Z_k = \theta \cdot \sqrt{I_k}$, where $I_k = 1/\sigma_k^2$ is called the information for analysis k , $k = 1, \dots, 3$, with σ_k^2 being the variance of θ_k . Hence, Z_k is a cumulative test statistic. Each of these statistics can potentially be calculated on a set of data, which we denote D_k . As the hypotheses are tested in sequence, we have that $D_1 \subset D_2 \subset D_3$. Figure 4.1 illustrates the accumulating data of such a trial. Figure 4.1 can also be used to illustrate a sequential meta-analysis. Suppose we have a meta-analysis with three planned trials. When one trial is finished, a hypothesis test is conducted. If the hypothesis is not rejected a second trial initiates and when finalised it is analysed together with the first trial in a meta-analysis. This scenario is illustrated in Figure 4.1 with the meta-analysis perspective written in parentheses.

Trials and meta-analyses might have very different reasons for having a sequential design. For trials, the main reason could be the expected reduction in the duration of the trial due to the chance of stopping early. It might also be for ethical reasons such that the patients get allocated to the most effective treatment as early as possible. The majority of sequential meta-analyses conducted do not have a pre-planned sequential design. However, for those that do, a reason might be that the individual trials are not expected to have sufficient power to reject the null hypothesis on their own. A sequential meta-analysis can then

combine the trials as they finish to potentially gain power while being able to reject the null hypothesis early at each update of the analysis. Early stopping of a sequential meta-analysis can reduce the combined cost of running the experiments as later trials do not need to be performed. Many meta-analyses are conducted without any central coordination of when or which new trials are set in motion. These decisions are taken by the local investigators. In these situations it is most appropriate to view the meta-analysis as retrospective. We will describe retrospective sequential meta-analyses in more detail at the end of this chapter. The reason for considering a sequential design in this set-up would often be to try to control the type-I-error when one believes that the meta-analysis has been updated at several times.

We will in the next section present the joint distribution of the testing sequence \mathbf{Z}_K . The joint distribution is used for creating the adjusted testing thresholds to ensure that the sequential trials type-I-error will not increase above the designed nominal level. Calculating testing thresholds is part of the software package developed which is presented in Manuscript II. The joint distribution can also be used for the calculation of point estimates which is the focus of Manuscript III. We will start with presenting how to calculate the testing thresholds before introducing how the joint distribution can be used for inference in Chapter 6.

Calculating testing thresholds in sequential designs

To control the probability of making a false-positive (type-I-error), we test a hypothesis at a specific test threshold related to the distribution of test statistic most often under the null hypothesis. In the case of only having one hypothesis test and assuming normality of the data, the test statistic could be the Z -score which will be compared to a threshold b for which the following holds:

$$P_{\theta=0}(|Z| > b) = \alpha.$$

Here α is the significance level and we test the null hypothesis $\mathcal{H}_0 : \theta = 0$ against the two-sided alternative $\mathcal{H}_A : \theta \neq 0$ where θ is the intervention effect of interest. We will in this thesis primarily consider two-sided tests. The software package has methods for both one-sided and two-sided tests.

We consider now a sequential testing scheme. Thus, testing the same null hypothesis more than once. Again using that $\mathcal{H}_0 : \theta = 0$ is the hypothesis of interest that we wish to reject, we let K be the number of times we plan to test the hypothesis. The sequence of test statistics is then $\mathbf{Z}_K = (Z_1, Z_2, \dots, Z_K)$ if we reach the final analysis. For the two-sided testing scenario we require a set of upper and lower test thresholds. The set of boundaries which we denote a_k for the lower boundaries and b_k for the upper are calculated to satisfy:

$$P_{\theta=0}(Z_k \leq a_k \text{ or } Z_k \geq b_k \text{ for some } k \in 1, \dots, K) = \alpha. \quad (4.1)$$

Figure 4.2 shows an example of a set of upper and lower stopping boundaries or testing thresholds, b_k and a_k respectively for $k = 1, \dots, 3$, which satisfy (4.1). Here the α level is set to 0.05 with expected interim analyses at 50% and 75% and the final analysis at 100% of the required information which we denote by timing on the x -axis of the figure. Figure 4.2 shows, via the green arrows, for which values of the test-statistics we will stop the trial due to rejection of

Timing	Sequential		Naive	
	Upper	Lower	Upper	Lower
0.5	2.96	-2.96	1.96	-1.96
0.75	2.36	-2.36	1.96	-1.96
1.0	2.01	2.01	1.96	-1.96
Type-I-error	0.05		0.097	

Table 4.1: Stopping boundaries and type-I-error for a sequential and naive testing scheme. The significance level was set to 5%.

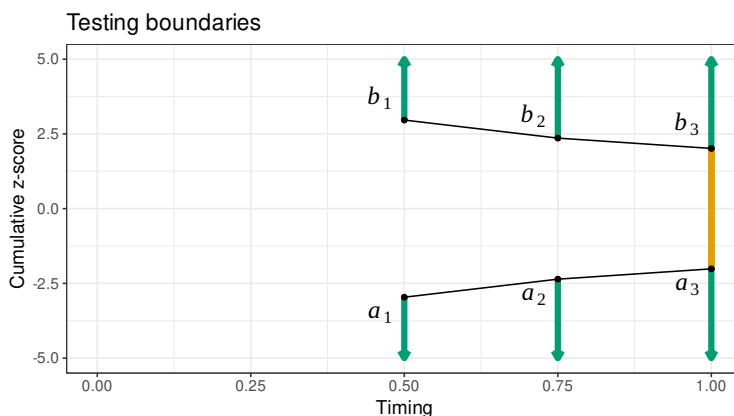


Figure 4.2: Stopping boundaries for a group sequential trial. The green arrows indicate which values of the test statistic at the given analysis we reject the null hypothesis. The yellow line indicate that we will stop the trial at the final analysis where the conclusion is that the null hypothesis could not be rejected.

the null hypothesis. The yellow line also represent when we will stop the trial, but without being able to reject the null hypothesis. Table 4.1 tabulates the boundaries from the figure comparing these with naive testing boundaries. The table also contains the type-I-error of both designs, where it is clear that the naive testing scheme will have an inflated type-I-error.

For now we have defined information to be the inverse of the variance of the estimate of intervention effect, but the information can also be depicted by the sample size. For information described via the inverse of the variance, a timing of 100% refers to the standard error reducing to the size for the trial to reach the designed level of power. The same holds for when information is described by sample size, a timing of 100% here translates to the sample size reaching the required size for achieving the power designed for in the trial. Given the potential multiple hypothesis tests, we can not allocate all of the type-I-error to the first hypothesis test. Instead we will split the probability of a false-positive between the planned interim analyses and the final analysis. This is known as α -spending. Splitting the α across the potential K analyses

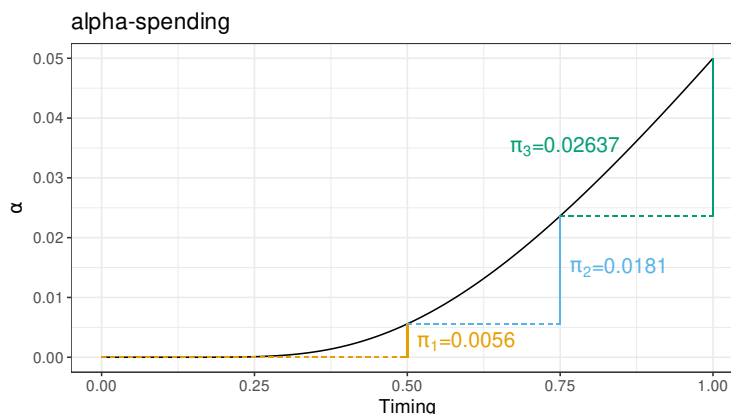


Figure 4.3: α -spending using the Lan and DeMets' version of O'Brien-Fleming stopping function with a α level of 0.05. Here the timing of the analyses are at 50%, 75% and 100% respectively.

can be described by:

$$\alpha = \sum_{k=1}^K \pi_k, \text{ where,} \quad (4.2)$$

$$\pi_k = P_{\theta=0}(a_1 < Z_1 < b_1, \dots, a_{k-1} < Z_{k-1} < b_{k-1} \text{ and } Z_k \leq a_k \text{ or } Z_k \geq b_k).$$

A visualisation of α -spending is shown on Figure 4.3. It shows how the type-I-error is split between three analyses planned at 50%, 75% and 100% of the timing. The number of analyses and their timings can be customised to the specific scenario at hand. The choice of number and analyses would typically depend on the expected behaviour of the data generating process and other factors.

To decide how much α is allocated at each analysis, one can use an error spending function. On Figure 4.3, the Lan and DeMets' version of the O'Brien-Fleming error spending function is used which uses a conservative amount of the type-I-error early in the trial to spend more at the end (Lan et al., 1983). The type-I-error spend at each analysis is shown on the figure, using the π notation from (4.2).

Distribution theory

The probability of crossing a boundary at the k^{th} analysis is dependent on all previous analyses not crossing their stopping boundaries. Thus, it is clear from the formulation of π_k in (4.2) that we are dealing with a joint probability distribution of the test statistics where we do not have independence between the potential K analysis times. Jennison et al. (1999) define a joint distribution of the test statistics, which apply to many common testing scenarios using

standardised test statistics:

$$\begin{aligned}
& \text{(i) } (Z_1, \dots, Z_K) \text{ is multivariate normal,} \\
& \text{(ii) } E(Z_k) = \theta\sqrt{I_k}, \\
& \text{(ii) } Cov(Z_k, Z_{k+1}) = \sqrt{I_k/I_{k+1}}.
\end{aligned} \tag{4.3}$$

We will use this multivariate distribution to calculate the boundaries $\mathbf{a}_K = (a_1, \dots, a_K)$ and $\mathbf{b}_K = (b_1, \dots, b_K)$. As the distribution of Z_k is dependent on Z_{k-1} , it makes the calculation of the boundaries cumbersome. Only the first test is easily defined where under the assumption of a normal distribution and having a two-sided test:

$$\begin{aligned}
Z_1 &\sim \mathcal{N}(\theta\sqrt{I_1}, 1) \quad \text{and} \\
\pi_1 &= \int_{-\infty}^{a_1} \phi(z_1 - \theta\sqrt{I_1}) dz_1 + \int_{b_1}^{\infty} \phi(z_1 - \theta\sqrt{I_1}) dz_1 \quad \text{where} \\
\pi_1/2 &= \int_{-\infty}^{a_1} \phi(z_1 - \theta\sqrt{I_1}) dz_1 \quad \text{and} \quad \pi_1/2 = \int_{b_1}^{\infty} \phi(z_1 - \theta\sqrt{I_1}) dz_1.
\end{aligned}$$

Here $\theta = 0$. Knowing the value of π_1 , one can calculate the boundaries a_1 and b_1 using the inverse cumulative distribution function of the normal distribution under the null hypothesis $\theta = 0$. However, when we consider $k > 1$, then the probability expression from (4.2) turns into a multivariate integral:

$$\pi_k = \int \int \cdots \int_{S_k} f_{k,\theta}(z_1, z_2, \dots, z_k) dz_k dz_{k-1} \dots dz_1.$$

Here $S_k = \{Z_j \in (a_j, b_j) \text{ for } j = 1, \dots, k-1 \text{ and } Z_k \leq a_k \text{ or } Z_k \geq b_k\}$ is the set of sample paths for \mathbf{Z}_k where the sequence of test statistics does not cross a boundary before analysis k and $f_{k,\theta}(z_1, z_2, \dots, z_k)$ denotes the joint density up to analysis k . Note that $f_{1,\theta}(z_1) = \phi(z_1 - \theta\sqrt{I_1})$ with ϕ denoting the standard normal distribution density function.

The joint distribution can be rephrased by utilising that Z_k is only dependent on Z_{k-1} , thus the Markov property holds for the sequence of test statistics. Using the distribution of the difference between two Z scores scaled by their information squared, we find that the conditional distribution of Z_k given Z_{k-1} is independent from Z_{k-2}, \dots, Z_1 . Let $\Delta_k = I_k - I_{k-1}$. Then:

$$Z_k\sqrt{I_k} - Z_{k-1}\sqrt{I_{k-1}} \sim \mathcal{N}(\theta\Delta_k, \Delta_k).$$

Deriving Z_k from the formula above, we can write the conditional distribution of Z_k given Z_{k-1} as:

$$Z_k|z_{k-1} \sim \mathcal{N}((\theta\Delta_k + z_{k-1}\sqrt{I_{k-1}})/\sqrt{I_k}, \Delta_k/I_k).$$

Which is an easier distribution to work with and we rewrite the multivariate integral as:

$$\pi_k = \int \int \cdots \int_{S_k} f_{1,\theta}(z_1) f_{2,\theta}(z_2|z_1) \dots f_{k,\theta}(z_k|z_{k-1}) dz_k dz_{k-1} \dots dz_1.$$

By using the recursive formula by Armitage et al. (1969) as presented in Jennison et al. (1999), we can instead for each analysis k compute the $k - 1$ previous integrals in turn. Let $f_{k,\theta}(z_k)$ be defined as:

$$f_{k,\theta}(z_k) = \begin{cases} f_{1,\theta}(z_1) & \text{for } k = 1 \\ \int_{a_{k-1}}^{b_{k-1}} f_{k-1,\theta}(z_{k-1}) f_{k,\theta}(z_k|z_{k-1}) dz_{k-1} & \text{for } k = 2, 3, \dots \end{cases}$$

Using the above formulation, we can calculate the boundaries a_k and b_k from:

$$\pi_k = \int_{a_k}^{b_k} f_{k,\theta}(z_k) dz_k. \quad (4.4)$$

This removes the multivariate integrals and changes the computation to consist of k successive single integrals. To show how it works, we look at $k = 2$, where we will use the conditional distribution of Z_2 :

$$f_{2,\theta}(z_2|z_1) = \frac{\sqrt{I_2}}{\sqrt{2\pi\Delta_2}} \exp\left(-\frac{(z_2\sqrt{I_2} - z_1\sqrt{I_1} - \theta\Delta_2)^2}{2\Delta_2}\right).$$

Under the two-sided test, we split the π_2 in two to solve for one boundary at a time. Here solving for b_2 , we get:

$$\begin{aligned} \pi_2/2 &= \int_{b_2}^{\infty} f_{2,\theta}(z_2) dz_2 \\ &= \int_{a_1}^{b_1} \int_{b_2}^{\infty} f_{1,\theta}(z_1) f_{2,\theta}(z_2|z_1) dz_2 dz_1 \\ &= \int_{a_1}^{b_1} f_{1,\theta}(z_1) \Phi\left(\frac{\theta\Delta_2 + z_1\sqrt{I_1} - b_2\sqrt{I_2}}{\sqrt{\Delta_2}}\right) dz_1, \end{aligned}$$

which is an easier expression to solve for b_k . The last line comes from calculating the last integral using symmetry of the cumulative distribution function of the normal distribution $\Phi(x)$, where $1 - \Phi(x) = \Phi(-x)$.

Raising the number of looks to k gives:

$$\begin{aligned} \pi_k/2 &= \int_{b_k}^{\infty} f_{k,\theta}(z_k) dz_k \\ &= \int_{a_{k-1}}^{b_{k-1}} \int_{b_k}^{\infty} f_{k-1,\theta}(z_{k-1}) f_{k,\theta}(z_k|z_{k-1}) dz_k dz_{k-1} \\ &= \int_{a_{k-1}}^{b_{k-1}} f_{k-1,\theta}(z_{k-1}) \Phi\left(\frac{\theta\Delta_k + z_{k-1}\sqrt{I_{k-1}} - b_k\sqrt{I_k}}{\sqrt{\Delta_k}}\right) dz_{k-1}. \end{aligned}$$

Hence calculating $f_{k,\theta}(z_k)$ can be done by calculating the previous $k - 1$ sub-densities starting from $f_{1,\theta}(z_1)$ and continuing to $f_{k,\theta}(z_k)$. This is how we calculate the stopping boundaries. Note that $f_{k,\theta}(z_k)$ for $k = 1, \dots, k$ are defined as sub-densities as each density is truncated by the stopping boundaries. Their sample paths are defined by the set S_k previously defined as $S_k = \{Z_j \in (a_j, b_j) \text{ for } j = 1, \dots, k - 1 \text{ and } Z_k \leq a_k \text{ or } Z_k \geq b_k\}$.

That $f_{k,\theta}(z_k)$ is a sub-density is seen from Figure 4.4 and 4.5. Both figures visualise the sub-densities at the first two interims of a group sequential trial

with three planned analyses at respectively 50%, 75% and 100% of the timing. Lan & DeMets' version of Pocock error spending is used in both plots with a type-I-error of 10% and a power of 90%. On Figure 4.4 is the intervention effect set to the null effect, hence $\theta = 0$. From the plots in Figure 4.4 it is shown that the sub-density for interim 2 is different from interim 1 due to the conditioning that one will only proceed to interim 2, if one did not stop in interim 1. Hence the statistics closer to the null effect will be parsed on to interim 2. This causes the sub-density of interim 2 to center more around the null effect as shown on the last plot on Figure 4.4.

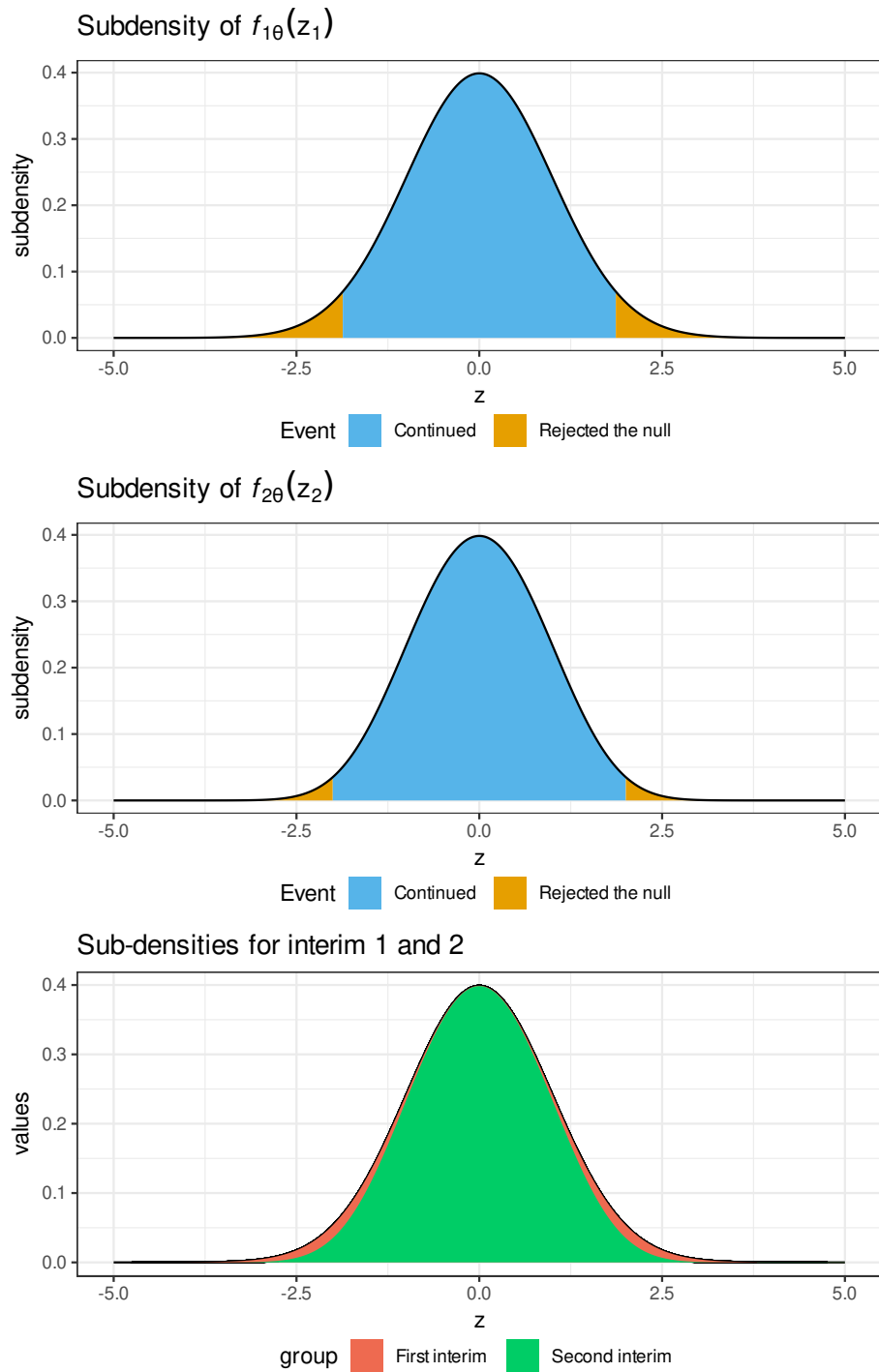
In Figure 4.5 there is a positive intervention effect, thus $\theta = \delta$, where δ is set such that we have 90% power. The reader should notice that when using Pocock-like error spending, the α -spend is constant over the interim analyses as a function of the timing. Thus, a large proportion of the statistics under $\theta = \delta$ will cross the first boundary. This is shown on the top plot Figure 4.5. The dependence of the first sub-density on the second sub-density is even more clear in this scenario and is visualised on the last plot of Figure 4.5. Here the sub-density of the second interim moves towards the null effect due to the selection of continued statistics from the first interim. This last plot is also a visualisation of why early stopped trials will tend to over-estimate the effect whereas trials continuing to the last analysis will tend to under-estimate the effect.

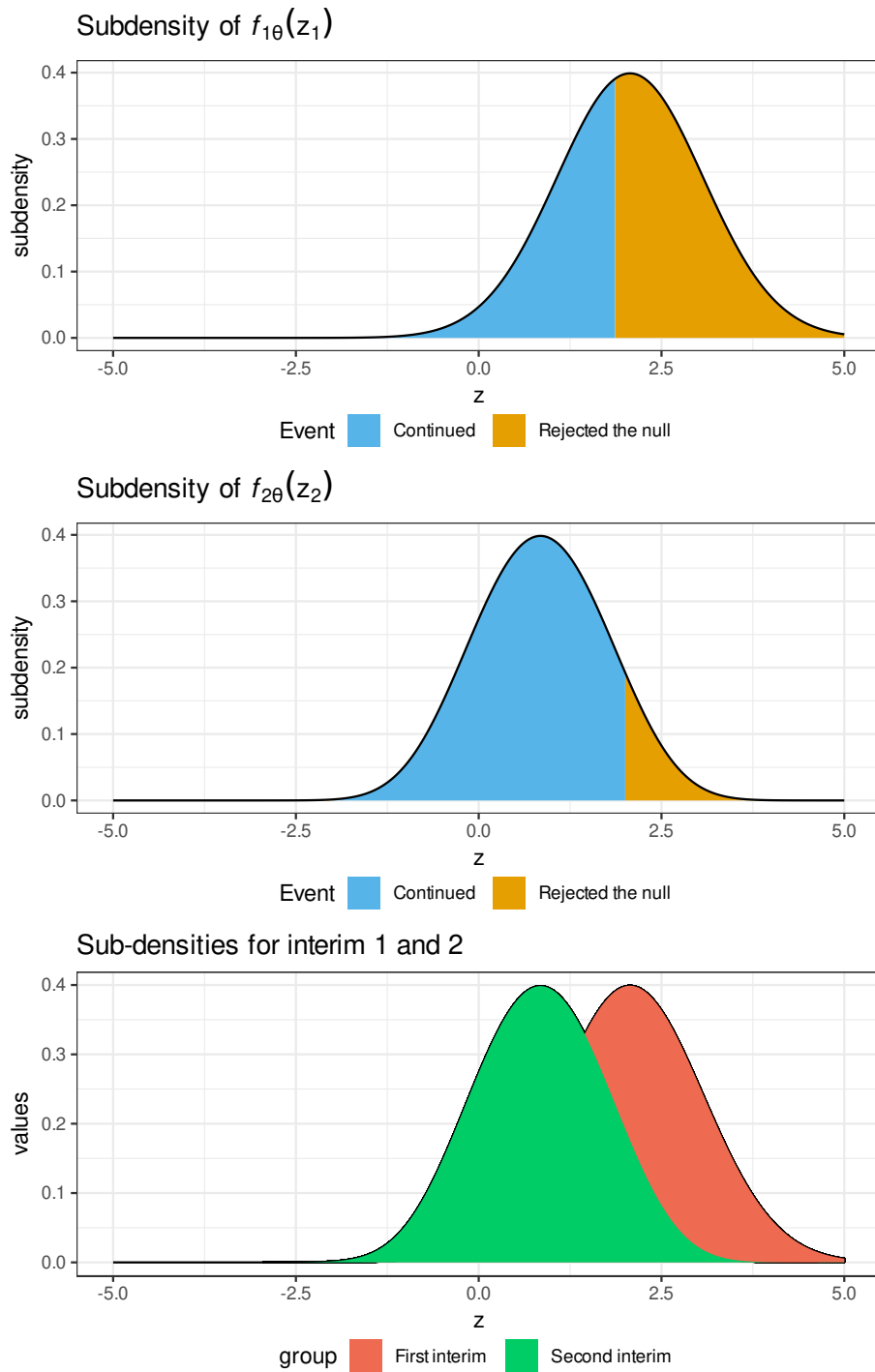
This was a rather long introduction to some of the theory behind group sequential methods. The section focused on type-I-error control in a group sequential design, but just as important is the type-II-error control which we will touch more upon in the next chapter. We will finish this chapter by how TSA uses group sequential methods for sequential meta-analysis.

4.3 Sequential meta-analysis using TSA

A sequential meta-analysis shares many of the same elements as a sequential clinical trial. But where the sequential trial is a very controlled process, a meta-analysis can be sequential in many ways with different levels of control of the process. The benefit from using TSA will differ depending on the scenario and control of the sequential meta-analysis. Different sequential meta-analysis scenarios include:

- Sequential retrospective meta-analysis:
 - Completely retrospective meta-analysis with no expected future updates, but it has been updated in its lifetime
 - Retrospective meta-analysis where trials could be added to be meta-analysis in the future
 - Retrospective meta-analysis where trials are planned be added to be meta-analysis in the future
 - Systematic living reviews
- Sequential prospective meta-analysis:
 - Prospective meta-analyses with a sequential design

Figure 4.4: Sub-densities of the first and second interim with $\theta = 0$

Figure 4.5: Sub-densities of the first and second interim with $\theta = \delta$

- Prospective meta-analyses without a sequential design but with planned updates

Note that the list only contains scenarios where there is either an expected or occurred update of the meta-analysis. We do not consider meta-analyses that will never be expected to or considered to be updated in the future in this thesis. The above list is split according to retrospective and prospective sequential meta-analyses, a characteristic which we will shortly describe. How well TSA is able to control for type-I- and type-II-errors is planned to be investigated for each of the scenarios presented above in future research.

Besides varying in the level of control, sequential meta-analyses also differ from sequential trials in terms of planning. An initial power/sample size calculation is always part of a sequential clinical trial, however, this is not always the case for sequential meta-analysis. A sample size or similar is necessary for using sequential methods as we need to define a definite stopping time and a definition of timings of the trials. For this purpose TSA uses *required information size* (RIS) and when heterogeneity is present it uses the diversity-adjusted required information size (DARIS) as the sample size which we consider to be the final analysis. Both RIS and DARIS were defined in Chapter 2. The timing of the sequential meta-analyses is described as the proportion of participants per meta-analysis out of the either RIS or DARIS in TSA. Using RIS or DARIS, TSA has a definition of the timing and can calculate the boundaries by setting the timing of the meta-analyses equal to the proportion of participants in the meta-analysis out of the required sample size.

The sample size of the sequential meta-analysis can be calculated before the conduct of the meta-analysis. This is the case for what is known as prospective meta-analyses where the design of the meta-analysis is set, and thus the sample size is calculated, prior the knowledge of any trial results (Seidler et al., 2019; Thomas et al., 2023). Most sequential meta-analyses are retrospective which means that the meta-analysis is calculated when already knowing the results of some of the trials included and a pre-calculated sample size is not available. The way TSA should be used is dependent on whether the meta-analysis is prospective or retrospective. This is described in Manuscript II. We will also see in Appendix B that the strength of the type-I-error control are different for prospective and retrospective meta-analyses and whether there is presence of heterogeneity or not.

When using sequential tests such as the ones we have just described, the guarantee for type-I-error control seems only to be satisfied for sequential analyses with complete control over the decision-making of whether the analysis stops or continues and if there is no heterogeneity (see Appendix B). This is not often possible with sequential meta-analysis. While being used for decision making regarding changes in standard of care for example, reaching a conclusion on the meta-analysis does not guarantee that the meta-analysis will not be updated at a later point. Similarly does concluding that the meta-analysis should continue not guarantee that new studies will be created. For this reason, most users of TSA will benefit from using TSA for the sequential meta-analysis as a sensitivity analysis. In this scenario TSA provides a perspective of the current results given the present information such as heterogeneity from a sequential test point of view. It can describe the current sequential meta-analysis' power and if an sequential test had been adopted from the beginning whether the

current results can be considered statistically significant. Given the complexity of sequential meta-analyses and its many potential versions, it might be of interest for the user to create different simulations using TSA to see the expected control of the type-I- and type-II-error given their unique data and meta-analysis process. The package developed which we present in Manuscript II can serve this.

We have in this section not touched much upon the topic of power. As mentioned in the introduction to group sequential methods, the sample size for sequential trials are often larger than the usual fixed-sample trial which is determined by an adjustment factor. This adjustment factor depends on the design of the sequential trial such as the chosen error spending function and the timing of the planned analyses. The TSA software implemented in Java does not adjust the sample size according to the sequential design. This has been updated in the R version of TSA which we will present in the next chapter: “Computational methods”. This chapter will introduce some of the computational challenges when translating and updating TSA to R which includes adjusting the sample size.

5. Computational methods

This chapter will introduce Manuscript II, an R (R Core Team, 2020) implementation of the Trial Sequential Analysis (TSA) software originally implemented in Java. As an introduction to the manuscript, we will start with a section concerning type-II-error and futility boundaries. This is followed by a section about some of the algorithms used in the software which concludes the introduction to Manuscript II.

5.1 Type-II-error in group sequential designs

We wish control the level of false-positives and false-negatives for a sequential analysis. Controlling type-I- and type-II-errors ensure that there is a small probability of falsely rejecting the null under the null being true, and a large probability for rejecting the null when the alternative is true. These criteria can be formulated as follows in a one-analysis scenario, where β is the type-II-error:

1. $P_{\theta=0}(|Z| > c) = \alpha$,
2. $P_{\theta=\delta}(|Z| > c) = 1 - \beta$.

Here we, to be consistent with the previous chapter, continue to consider two-sided tests. We extended the first criteria in the last chapter to sequential trials and meta-analyses where we are exposed to potentially a sequence of test statistics \mathbf{Z}_K where K is the planned number of analyses. Now, we will focus on the latter criteria, concerning the power of the analysis and extend it to a sequential testing scheme. Here, and for the code in the R implementation of TSA, we will for all practical purposes have that:

$$\begin{aligned} P_{\theta=\delta}(|Z| > c) &= P_{\theta=\delta}(Z > c) = 1 - \beta, \quad \text{and,} \\ P_{\theta=-\delta}(|Z| > c) &= P_{\theta=-\delta}(Z < -c) = 1 - \beta. \end{aligned}$$

Thus, we assume that $P_{\theta=-\delta}(Z > c) = P_{\theta=\delta}(Z < -c) = 0$ as done in Jennison et al. (1999).

In the sequential set-up we have the following expression of the power:

$$\sum_{k=1}^K P_{\theta=\delta}(a_i < Z_i < b_i \text{ for } i = 1, \dots, k-1 \text{ and } Z_k > b_k). \quad (5.1)$$

Here the set of upper and lower boundaries, \mathbf{a}_K and \mathbf{b}_K , have already been calculated, when controlling for the type-I-error. Note that the expression in (5.1) will most likely not equal the desired level of power, $(1 - \beta)$. To

get to the desired level, we adjust the information by an adjustment factor R . This is always done in sequential trials to get the correct power. The adjustment factor adjusts the timings of the analyses by scaling the information from $\mathbf{I}_K = (I_1, \dots, I_K)$ to $R \cdot \mathbf{I}_K$. How it works can be shown via an example. Consider the power for a sequential design with two planned analyses. Then the adjustment factor will ensure that the sequential trial is well powered by solving the following equation for a given value of β :

$$\begin{aligned} \sum_{k=1}^2 P_{\theta=\delta}(a_i < Z_i < b_i \text{ for } i = 1, \dots, k-1 \text{ and } Z_k > b_k, R) &= 1 - \beta, \\ P_{\theta=\delta}(Z_1 > b_1, R) + P_{\theta=\delta}(Z_2 > b_2 | Z_1 < b_1, R) &= 1 - \beta, \\ \int_{b_1}^{\infty} f_{1,\theta=\delta}(z_1, R) dz_1 + \int_{b_2}^{\infty} f_{2,\theta=\delta}(z_2, R) dz_2 &= 1 - \beta, \\ &\Phi(z_1 - \theta\sqrt{R \cdot I_1}) + \\ \int_{a_1}^{b_1} f_{1,\theta=\delta}(z_1, R) \Phi\left(\frac{\theta\Delta_2 \cdot R + z_1\sqrt{I_1 \cdot R} - b_2\sqrt{I_2 \cdot R}}{\sqrt{\Delta_2 \cdot R}}\right) dz_1 &= 1 - \beta. \end{aligned}$$

Using the adjustment factor changes to the sample size requirement. The factor's size will depend on the number of tests, the error spending function, and the timing of the analyses, increasing as the number of planned analysis increases. The adjustment factor is part of the R implementation of Trial Sequential Analysis (TSA), but was not included in the original TSA software. In the scenario of a prospective sequential meta-analysis with a formal sequential design, it makes sense that we would want our analysis to have the right power. Hence, an adjustment factor R is in this scenario needed. In other scenarios, where TSA is used more as a perspective or sensitivity analysis rather than a primary design, it is still of interest to have the perspective of a sequential design where both type-I-error and type-II-error are projected. The interpretation and use of TSA under the different sequential meta-analyses scenarios are discussed in Chapter 7.

Note that for trials with only a few analyses, the adjustment factor will be close to 1. However, this changes when futility boundaries are introduced. Another feature which is part of the RTSA package, but not included in the original software, is binding futility boundaries. Futility boundaries will be introduced in the next section.

Futility boundaries

So far we have defined boundaries created for rejecting the null hypothesis in a sequential test. Another common set of boundaries to consider are futility boundaries. This type of boundary defines the values in the sample paths of \mathbf{Z}_K where continuing the analysis for rejection of the null hypothesis is futile. Figure 5.1 shows two examples of futility boundaries together with the boundaries for rejecting the null hypothesis. As before the yellow lines define for which values we would stop the analysis with the conclusion that the null hypothesis could not be rejected. We define the boundaries $\mathbf{d}_K = (d_1, \dots, d_K)$ and $\mathbf{c}_K = (c_1, \dots, c_K)$ to be respectively the upper and lower futility boundaries in a two-sided design. For a one-sided design, only \mathbf{d}_K will

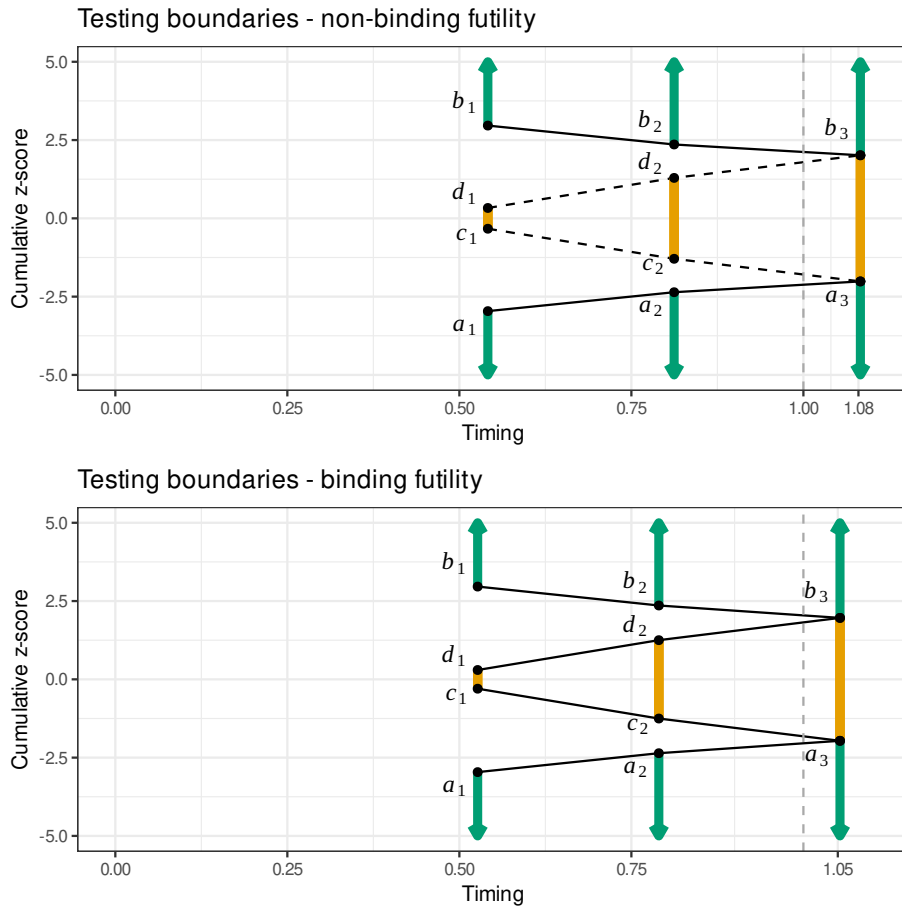


Figure 5.1: Group sequential designs with futility boundaries. The upper plot visualises non-binding futility boundaries shown with dashed lines.

be calculated. Figure 5.1 also shows, indirectly, the adjustment factors for the designs on the x -axis. For the upper plot an adjustment factor of $R = 1.08$ is required. Thus, 8% more information is required compared to a similar analysis designed with only one analysis planned. How strict these futility stopping boundaries should be enforced depends on whether one chooses to use non-binding or binding futility boundaries. We start with defining the non-binding futility boundaries.

Using non-binding futility boundaries, entering the futility area does not imply that the analysis necessarily needs to stop. Figure 5.1 upper plot is a sequential design with non-binding futility boundaries. This means that the trialist or meta-analyst can decide whether to continue to the next planned analysis or stop. Under such a flexible testing regime it is still of interest to ensure that the type-I-error does not increase and that the power does not decrease. This is achieved by first setting the control of the type-I-error as if one will never stop for futility. Then, the control of the type-II-error, adjusted by R , should be set as if one would always stop for futility. These two steps

	Not stopping for futility	Always stopping for futility
Type-I-error	5%	4.6%
Type-II-error	8%	10%

Table 5.1: Probability of type-I- and type-II-errors in a sequential design with non-binding futility boundaries. The design was based on a type-I-error of 5 % and a type-II-error of 10 % and three planned analysis at 50%, 75% and 100% of the timing.

keep the α -spending stopping boundaries at the same level as in the scenario with no futility boundaries. However, the adjustment factor R will be larger than the design without futility boundaries. The reason being that if the trialist will always stop for futility, we need a larger sample size. Table 5.1 shows the control of the type-I- type-II-error given the two extremes that one would always stop or never stop if entering the futility area when the futility boundaries are non-binding. In each case the probability of these types of errors is less than or equal to the nominal set levels.

Futility boundaries can also be designed as binding. This means that when the test statistic enters the futility area, the analysis must be stopped. Binding futility changes the null rejection stopping boundaries, hence the boundaries \mathbf{b}_K and \mathbf{a}_K change. The reason is that the risk of making a type-I-error decreases as one will stop for futility if the futility area is entered. Thus the null rejection stopping boundaries will reduce compared to the design without futility boundaries and the design with non-binding futility boundaries. Figure 5.1 lower plot shows a design with binding futility boundaries. Besides reducing the testing thresholds for the null rejection boundaries, binding futility designs also requires a smaller sample size compared to the non-binding futility design, which is seen by comparing the x -axes of the two plots in Figure 5.1. The requirement of the analysis to stop when reaching the futility area might be too restrictive for retrospective meta-analyses where there is no guarantee that new studies are not initiated and later added to the meta-analysis.

The calculation of the futility boundaries are closely related to the calculation of the null rejecting boundaries. As stopping the analysis for futility is based on the type-II-error instead of type-I-error, futility boundaries are often called β -spending boundaries whereas the null rejection boundaries are called α -spending boundaries. In the same way as α -splitting, presented in the previous chapter, we can split the β across the planned analyses. Thus,

$$\beta = \sum_{k=1}^K \pi_k^\beta \quad \text{where,}$$

$$\pi_k^\beta = P_{\theta=\delta}(d_1 < Z_1 < b_1, \dots, d_k < Z_k < b_k).$$

For the α -splitting, if we have a two-sided test, the value of π_k was split in two. This is not the case for the β -spending. For the two-sided design will the futility boundaries be calculated such that there is $1 - \beta$ level probability of being above \mathbf{b}_K for $\theta = \delta$ and $1 - \beta$ level probability of being below \mathbf{a}_K for $\theta = -\delta$. To have a cohesive test, the last β spending boundaries must equal the last α spending boundaries, hence $c_K = a_K$ and $d_K = b_K$ in the two-sided design. Thus, the final analysis K will always be conclusive. Computing the

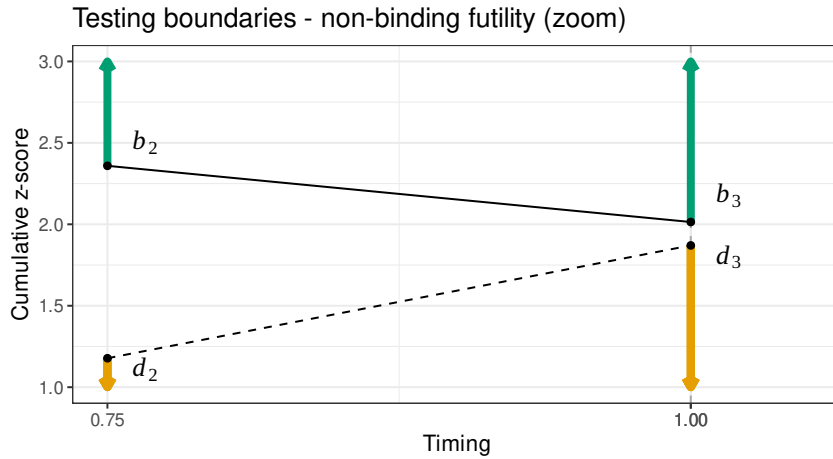


Figure 5.2: Group sequential designs with futility boundaries. If the information is not adjusted, the futility boundaries will not meet the α -spending boundaries.

boundaries naively will result in $d_K \neq b_K$ as shown in Figure 5.2. How to achieve $b_K = d_K$ is described in the Computational methods section coming up.

The reader should note that the calculation and programming of two-sided binding futility boundaries was a contribution by this thesis. While there was no R packages for sequential meta-analysis before RTSA, there exist R packages for group sequential methods for trials such as the `gsDesign` package (Anderson, 2022) and the `RPACT` package (Wassmer et al., 2022) which include designs with similar stopping boundaries. However, either these do not include two-sided binding futility boundaries or the functions are under development. One reason for this is that group sequential trials are often one-sided and it is becoming the norm to use non-binding futility boundaries. However, used for meta-analysis, it might be of interest to have binding futility boundaries as the role of the meta-analysis is often to illustrate the current evidence.

5.2 Computational methods

For now we have presented primarily theoretical results. This section is going to describe how some of the methods presented were practically implemented in the RTSA package (Soerensen et al., 2023b). As the package consists of approximately 5500 lines of code, we are going to spare the readers and only present a short summary of the key elements of the implementation.

Numerical integration

The calculation of the boundaries, both α and β spending required evaluating integrals of sub-densities. There are many different ways to solve integrals numerically. In the original TSA software, the trapezoidal rule for calculating the integrals was used. In the RTSA implementation, this changed to Simpson's

rule as described in Jennison et al. (1999). There were two reasons for changing from the trapezoidal rule to Simpson's rule. The first being that other boundary calculating packages in R used Simpson's rule and it would then be easier to compare the results of the RTSA package to the other packages for error/quality control. The second reason is that the method is based on quadratic interpolation which is known to generally perform better than the trapezoidal rule which is based on linear approximation. The numerical integration is programmed in C++ to reduce the computation time.

Algorithm 1: Boundary and adjustment factor calculation for a two-sided designs

initialisation

input: Specify α , β , whether the test is one- or two-sided, error spending functions for α - and potentially β -spending, choice of futility boundaries (none, non-binding or binding), and planned timings of the analyses.

begin

Based on the input calculate the α -spending boundaries \mathbf{b}_K and

\mathbf{a}_K . **if** *futility is set to none* **then**

| Calculate R based on \mathbf{b}_K , \mathbf{a}_K , β and \mathbf{I}_K

else

if *futility is set to non-binding* **then**

1. Calculate \mathbf{d}_K

2. Warp the information such that $d_K = b_K$

3. Remove boundaries where $\mathbf{d}_K < 0$ and re-calculate \mathbf{d}_K

4. Warp the information such that $d_K = b_K$

5. Using the warping factor as R , re-calculate \mathbf{d}_K and set $-\mathbf{d}_K = \mathbf{c}_K$

if *futility is set to binding* **then**

1. Calculate \mathbf{d}_K

2. Re-calculate \mathbf{b}_K conditional on \mathbf{d}_K

3. Warp the information such that $d_K = b_K$

4. Remove boundaries where $\mathbf{d}_K < 0$ and re-calculate \mathbf{d}_K

5. Warp the information such that $d_K = b_K$

6. Using the warping factor as R , re-calculate \mathbf{d}_K

7. Calculate the estimated type-I-error $\hat{\alpha}$

while $|\hat{\alpha} - \alpha| > tol$ **do**

| 7.a. Redo step 1 to 7 to update \mathbf{b}_K , \mathbf{d}_K and $\hat{\alpha}$

8. Set the warping factor to R and set $\mathbf{c}_K = -\mathbf{d}_K$ and

$\mathbf{a}_K = -\mathbf{b}_K$

Return \mathbf{a}_K , \mathbf{b}_K , \mathbf{c}_K , \mathbf{d}_K and R

Complete design

To derive a complete design with both α - and β -spending boundaries with both the type-I- and type-II-error under control, a specific ordering of the calculations is required. We show this ordering in Algorithm 1 describing a condensed version of how the two-sided boundaries with and without futility

boundaries are calculated. Just as in the scenario with no futility boundaries, we need to increase the sample size for reaching the right power. By increasing the timings of the designs, we are also able to make the boundaries meet as increasing the sample size will push the futility boundaries away from the null and closer to the α -spending boundaries. Setting the last futility boundary equal to the last futility α -spending warps the information scale. This warp scale is equal to the adjustment factor. The full algorithm can be seen above.

Testing

Like any software implementation, testing was the largest part of the work. Initially the goal of the testing was to ensure that the R implementation of TSA was in fact equal to the original TSA implementation. The first implementation started by using the TSA manual (Thorlund et al., 2011), where methods for meta-analysis and sample size calculation was coded from scratch but methods for calculating the boundaries used existing software in R, specifically the `gsDesign` (Anderson, 2022) package. The result of several meta-analyses analysed using TSA was then compared to the R implementation. Here it was found early on that the R implementation did not match the original software in terms of the boundaries, while the meta-analysis result was identical.

To resolve this issue of the mismatch between the boundaries, the solution ended up being a direct translation of the boundary calculating methods in TSA from Java to R. Hence the Java source code was translated to R and C++. This provided complete clarity of the methods used in TSA and it was possible to perfectly match the original software using circa 10 examples both the meta-analysis results and calculated boundaries. The translation also verified the following:

- The TSA futility boundaries were non-binding futility boundaries which was not clear until the translation.
- Sample sizes are only adjusted according to presence of heterogeneity but not by the sequential design.
- One-sided futility boundaries are implemented as two-sided futility boundaries.

Based on the findings, it was decided that the R implementation should, rather than replicate the original software, try to implement an updated version which followed the methods described in Jennison et al. (1999) chapter 19. The methods was implemented without the usage of other boundary calculating packages in R. However as some of these packages uses the same methodology for the calculation of boundaries, the boundary calculations where tested against `gsDesign` (Anderson, 2022) and `RPACT` (Wassmer et al., 2022).

5.3 Final remarks on the package

This thesis has been mostly concerned with the boundary calculations implemented in the `RTSA` package. However, as the original TSA software it also contains methods for meta-analysis and sample size estimation. These methods

were extended in the RTSA implementation to include a larger library of methods. Most of the methods were described in the documentation such as in the RTSA manual (see Appendix A) and via the package vignettes (<https://cran.r-project.org/web/packages/RTSA/index.html>).

The original objective for the implementation of the package was to be able to compare TSA to other sequential methods in terms of type-I- and type-II-error control. While this is part of future work some initial results concerning only TSA is presented in Appendix B.

Note that the package also includes methods for calculating inference. Due to the sequential structure, naively calculated estimates, confidence intervals and p-values are invalid. The RTSA package has implemented TSA-adjusted confidence intervals, stage-wise adjusted confidence intervals and p-values, and the median unbiased estimator. For information about these methods, see the next Chapter 6. With these final remarks we conclude the introduction to Manuscript II. The next chapter will introduce Manuscript III.

6. Sequential conditional bias

The previous two chapters have focused on the design of a sequential testing regime. This chapter will focus on the analysis of sequential data. We will both consider sequential data with and without a formal sequential testing regime. We will focus on how to carry out inference which we consider here to be point estimates, confidence intervals, and p-values. The chapter will end with a section concerning conditional bias due to decision making, a type of bias which can happen in sequential meta-analysis. This topic is the focus of Manuscript III.

6.1 Context

Before we present the background for Manuscript III, we specify what kind of updated meta-analyses we will be considering later in this chapter and in Manuscript III. We are the context of updated meta-analyses which we also call sequential meta-analysis here with no formal requirement of a sequential testing regime. This means that we consider analyses (trials or meta-analyses) that are updated with new trials in their lifetime (a meta-analysis). Specifically, we investigate scenarios where the new trials are motivated by the result of a previous underpowered meta-analysis or trial from which no conclusion was drawn but the initial results are promising. Here promising refers to the observed test statistic of the previous analysis being in the direction of benefit of intervention. We are especially interested in the point estimate in the updated meta-analysis under this decision scheme, as the considered continuation mechanism can result in a type of sequential conditional bias of the point estimate of the updated meta-analysis.

Group sequential trials have created adjustment estimators to adjust for the effect the group sequential design on the point estimate in a group sequential trial. We will in the following section describe how the bias occur in a group sequential trial along with different solutions to this problem. The chapter ends with an introduction to how we might coerce some of the solutions from the group sequential trial setting to our scenario, the updated meta-analysis. It further includes one of the motivations for the research in Manuscript III; the conclusive updated meta-analysis for investigating the effect of early versus delayed cord-clamping on pre-term delivery babies' risk of in-hospital mortality (Fogarty et al., 2018).

6.2 Analysis of sequential data for sequential trials

Consider the scenario where we have a continuously updated trial where we stop when significance is shown or when we reach a sample size for which we achieve the wanted power. When no group sequential testing scheme is used, we may have biased and invalid inference if the inference is calculated naively. Here invalid means that the usual interpretation of the p-value is invalid and confidence intervals do not have desired coverage probability. That the interpretation of the p-value is invalid was seen in Chapter 4 where it was discussed how repeating a hypothesis test will increase the type-I-error above the nominal level, see Table 4.1 for an example. Thus, when analysing sequential data naively without a sequential testing scheme, the interpretation of naively computed p-values is invalid. Due to the relationship between confidence intervals and p-values, the interpretation of the naive confidence interval will also be invalid (Armitage et al., 1969; Jennison et al., 1984). Also, if the continuation of the analysis depends on whether the test statistic crosses a testing threshold, the point estimate will be biased. This is because the sampling path of the point estimate of sequential analyses is not normally distributed conditional on crossing a testing threshold. This was visualised in Figure 4.5 for the test statistic, which is highly related to the size of point estimate. Hence for analysing sequential data, the naively calculated metrics of inference are most likely invalid and biased.

Still considering the scenario where we have sequential data, but now using a group sequential testing scheme, we may still have biased and invalid inference if the inference is calculated naively. While the type-I-error, in connection to the decision making of whether the null hypothesis is rejected, is controlled to its nominal level, the naively calculated p-value itself is still invalid. This is again due to the non-normal sampling path of the test statistics. This is clear from the calculation of the p-value. Consider the usual definition of the p-value for the two-sided design:

$$2 \cdot P_{\theta=0}(Z \geq z),$$

where z is the observed test statistic and Z is the Z -score in the standard normal distribution. This definition does not take into account the sequential testing sample path. For this reason another definition has been used in the sequential setting:

$$2 \cdot P_{\theta=0}((Z, t^*) \geq (z, t)), \quad (6.1)$$

where we now are considering a set containing the Z -score but also the actual stopping time t . Here t^* defines the optional stopping times. Now considering a set instead of a single value, we will need to define rules for which $(Z, t^*) \geq (z, t)$. There is a variety of such orderings. We will here only present what is known as the stage-wise ordering (Armitage, 1957). Here we have that $(Z, t^*) \geq (z, t)$ for:

- $t^* = t$ and $Z \geq z$,
- $t^* < t$ and $Z \geq b_{\{t^* < t\}}$,
- $t^* > t$ and $z \leq a_{\{t^* > t\}}$.

Suppose a sequential analysis stops at the second interim ($t = 2$) crossing the upper stopping boundary b_2 with the observed test statistic z_2 . In this scenario the following sets will be taken as more extreme following the stage-wise ordering:

- $t^* = 2$ and $Z \geq z_2$,
- $t^* < 2$ and $Z \geq b_1$.

The last general rule does not apply in this scenario since $z_2 > a_{\{t^* > t\}}$ for $t = 2$. As we can compute the probabilities for each element of the ordering, we can now compute p-values, confidence intervals and point estimates which are valid. Starting with the p-value, we have already defined the upper p-value in (6.1). By upper, we refer to the case where the upper boundary was crossed. In the scenario of crossing the lower boundary, the same p-value can be used for the two-sided test due to symmetry. To compute the confidence intervals, we can invert the hypothesis tests such that:

$$P_{\theta_U}((Z, t^*) \geq (z, t)) = P_{\theta_L}((Z, t^*) \leq (z, t)) = \alpha/2,$$

using the notation from Jennison et al. (1999), where θ_L and θ_U are respectively the lower and upper interval limits for the intervention effect θ . Thus we also have a way to calculate the confidence limits. A similar computation can be done to find the median unbiased estimate, denoted θ^* below, which is found via:

$$P_{\theta^*}((Z, t^*) \geq (z, t)) = 0.5.$$

Stage-wise confidence intervals, p-values, and the median-unbiased point estimate are implemented in the RTSA package (Soerensen et al., 2023b). There are other types of orderings to consider when computing the inference. These and the stage-wise ordering can only be computed when the analysis stops, thus when a boundary is crossed at interim or at end-of-trial.

In Jennison et al. (1984) the goal was to provide an interval which could be used at any stage in the sequential design and not just when crossing a stopping boundary. They created another way to compute intervals for the point estimate using what is known as repeated confidence intervals (Jennison et al., 1984). The repeated confidence intervals are created such that the coverage is guaranteed for the sequence of confidence intervals:

$$P_{\theta}(\theta \in I_k \text{ for all } k = 1, \dots, K) = 1 - \alpha \quad \text{for all } \theta.$$

The guarantee for the sequence affects the coverage of the single interval. The single interval will be wider, thus more conservative. The repeated confidence intervals are implemented in the RTSA package (Soerensen et al., 2023b) but described as TSA-adjusted confidence intervals to be consistent with the naming in the original TSA software (Thorlund et al., 2011).

Conditional point estimates

We will now focus on the point estimate in a sequential analysis which is the topic in Manuscript III. As we will be focused on the bias of the point estimate, theory about the distribution of the point estimate is required.

As in Jennison et al. (1999), the sample density of the point estimate is defined as:

$$\sum_{k=1}^K f_{k,\theta}(z_k) \sqrt{I_k}, \quad \text{with expectation,} \quad E_{\theta}(\hat{\theta}) = \sum_{k=1}^K \int_{S_k} \frac{z}{\sqrt{I_k}} f_{k,\theta}(z_k) dz,$$

where S_k defines the continuation region for Z_1, \dots, Z_{k-1} and that $Z_k \notin (a_k, b_k)$ defining at what stage the analysis stopped. Consider the naively calculated point estimate which we denote $\hat{\theta}$. From the above section we know that this estimate is biased. To adjust the point estimate for bias, one option is to calculate the marginal or unconditional bias adjustment which is defined as:

$$\begin{aligned} \tilde{\theta} &= \hat{\theta} - b(\hat{\theta}) \\ b(\hat{\theta}) &= E_{\theta}(\hat{\theta}) - \theta. \end{aligned}$$

Here $b(\hat{\theta})$ is the bias of the observed point estimate. As it is not possible to know the true value θ , the adjusted estimate $\tilde{\theta}$ is calculated via solving the following equation:

$$0 = E_{\tilde{\theta}}(\hat{\theta}) - \tilde{\theta},$$

where $\tilde{\theta}$ is then the adjusted estimate of θ . Notice that the stage-wise adjusted point estimator is another type of unconditional estimator of the intervention effect.

Now, the estimator just presented does not condition on the stopping time, which is known to influence the size and direction of the bias (Fan et al., 2004). To adjust for the stopping stage, conditional estimators can be used which will solve:

$$0 = E_{\bar{\theta}}(\hat{\theta}|t^*) - \bar{\theta}. \quad (6.2)$$

For this case, it has been found that $\bar{\theta}$ can be estimated by maximising the conditional log-likelihood (Liu et al., 2004):

$$\begin{aligned} l_{k,\theta}(z_1, \dots, z_k|t = k) &= \log(f_{k,\theta}(z_1, \dots, z_k|t = k)) \\ &= \log(f_{k,\theta}(z_1, \dots, z_k)) - \log(P_{\theta}(t = k)). \end{aligned}$$

Thus, $\bar{\theta} = \max_{\theta} l_{k,\theta}(z_1, \dots, z_k|t = k)$ as $\bar{\theta} = \hat{\theta} - \int_{S_k} (\hat{\theta} - \theta) P_{\theta}(t = k) d\theta$. Using the conditional estimator, the adjusted point estimate will be conditionally unbiased. The conditional estimator is used in Manuscript III.

To summarise the conclusions so far, the point estimate will be biased in a group sequential testing regime as a consequence of the stopping boundaries which is a purely design-affected bias if calculated naively. The conditional estimator adjusts the point estimate to provide a conditional unbiased estimate. Thus, we have an unbiased estimate in a sequential analysis, if the decision to continue the analysis strictly follows the design. The focus in Manuscript III is a scenario where this is not the case. Shifting our attention to bias stemming from the data used in the analysis is presented in the next section. Before this, we comment that this section only looked at inference expressed via point estimates, confidence intervals, and p-values. There are other types

of inference that can be computed, which include Bayesian metrics of inference such as credible intervals (Spence et al., 2016) or inference arising from game-theory such as e-values and anytime-valid confidence intervals computed from e-variables (ter Schure et al., 2019). Comparison with these types of inferential measures was not part of this thesis.

6.3 Conditional bias due to decision making

When deciding whether to conduct new research it is common to look at similar earlier conducted trials for evidence that one’s research ideas are worth investigating and that additional research is required. Based on the evidence, the decision to conduct a new trial can be based on whether the evidence looks promising, such as pointing in the direction of interest or the point estimate might be larger or equal to than some minimal clinical difference. If the evidence is insignificant, more information may be needed and may justify new trials. This was the case for the Australian Placental Transfusion Study (APTS) (Tarnow-Mordi et al., 2017) and the LIFT study (Tarnow-Mordi et al., 2020). Both trials were motivated by pre-existing meta-analyses and were then combined in each their meta-analysis with the evidence used for justifying them. This type of dependence on the results of earlier evidence can lead to bias if the new trial will be combined with the earlier evidence (Kulinskaya et al., 2015). It can be thought of as a kind of selection bias. While the new trial will not be biased, the cumulative set of trials (old and new) might be since the decision to “continue” the analysis was dependent on the sign of the test statistic or the size of the point estimate from the previous trials. Figure 6.1 shows an illustration of this problem under the assumption of no true intervention effect. In the figure, the top plot illustrates a distribution of point estimates under the null hypothesis and decision rules stating when the analysis will continue with the addition of a new trial based on a range of values of the observed z-score. The bottom plot illustrates the point estimate of the continued analysis where the old and new information is combined. Here the new information is also simulated under the null hypothesis being true. From this example we see that the point estimate will be skewed towards a larger effect based on the decision rules. This, how the second analysis becomes biased due to the decision to continue based on promising but non-significant earlier results. Furthermore, it is important to note that the first analysis was not biased, it was the selection mechanism that creates the bias.

The size of this problem has been previously investigated. Kulinskaya et al. (2015) investigate the bias of the point estimate of an updated meta-analysis, where the probability to continue a meta-analysis is positively correlated with the size of the earlier point estimate. This was found to lead to an increase in bias in the updated meta-analysis. They named the scenario sequential decision bias and emphasised the problem by investigating the bias as a function of the size of the earlier point estimate. The paper did however not provide any solutions to the problem. ter Schure et al. (2019) also look into the this problem and defines a term called Accumulation Bias. Similar to how we view the problem, ter Schure states accumulation bias to be a result of: “some studies or meta-analyses not being performed at all, as a result of previous findings in a series of studies.” (ter Schure et al., 2019). Which is the same

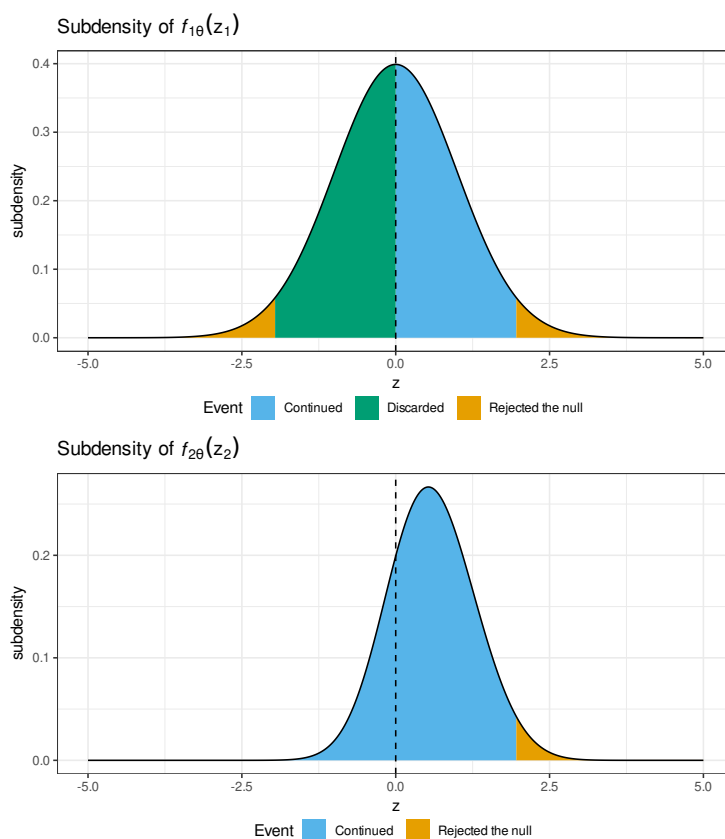


Figure 6.1: Visualisation of the potential impact of sequential selection bias when only promising non-significant studies are continued and updated with new information. The dashed line represent the true value of the effect estimate which is 0.

problem as we identify but seen from the perspective of the trials not continued due to some decision process stemming from the results from current evidence. ter Schure et al. (2019) do not focus on adjusting the bias it rather recommends to adopt another paradigm named the “ALL-IN” approach. The research goal of Manuscript III is specific to the adjustment of the point estimate and we want to see whether we can adjust for the bias without changing to a new statistical paradigm. The method we will propose in Manuscript III will create an unbiased estimate in a sequential meta-analysis using an adapted version of the conditional estimators just presented for group sequential trials. To introduce Manuscript III we will look at one of the motivations for the paper in more depth and using that example provide a more formal definition of the set-up and scenarios we are investigating a solution to.

Motivation and set-up for Manuscript III

Early versus delayed cord clamping

In 2012, a Cochrane review was released investigating the effect of delayed versus early cord clamping on in-hospital death in pre-term newborns. The review showed an in-significant reduction in in-hospital mortality, favouring the delayed cord clamping over early cord clamping (Rabe et al., 2012). Based especially on this review, a justification paper was published for the Australian Placental Transfusion Study (APTS) a large randomised controlled trial investigating the effect of delayed vs early cord clamping on pre-term newborns mortality (Tarnow-Mordi et al., 2014; Tarnow-Mordi et al., 2017). The study did not manage to prove a significant reduction in in-hospital mortality but did show promising results of delayed cord clamping over early cord clamping as the Cochrane review did.

A later meta-analysis also had the aim of evaluating the effect of delayed cord clamping vs early cord clamping (Fogarty et al., 2018). It combined the two sources of evidence by updating the Cochrane review with APTS (Fogarty et al., 2018). Combining the two showed a significant reduction in in-hospital mortality of pre-term newborns under the intervention of delayed cord clamping versus early cord clamping. The result was a relative risk of 0.71 with 95% CI (0.53; 0.95) and a p-value of 0.02. This result could be influenced by the decision to continue the Cochrane review based on its promising results and that it was underpowered (Tarnow-Mordi et al., 2014). While the result of APTS itself is unbiased, the decision to create it was influenced by the results of the Cochrane review. Had the Cochrane review not shown a tendency towards benefit of delayed cord clamping (the intervention), it seems unlikely that APTS would have been initiated as it would be more likely that one wanted to stick to the standard-of-care (early to no cord clamping). Furthermore, as it was also noted in the justification paper of APTS, the decision to contribute with a large RCT was also based on the power/imprecision of the Cochrane review (Tarnow-Mordi et al., 2017). This means that had the conclusion of the Cochrane review been significant, the initiation of APTS was not considered as needed.

To update a meta-analysis simply because it lacks power does not induce bias of the point estimate in the updated meta-analysis. However, this is not true under a conditional updating scheme, where an update would only have been initiated due the observed results pointing in a certain direction. This does not mean that one should necessarily remove the early evidence from the meta-analysis. If there is no concern about the quality of the earlier evidence, it might be difficult to argue that one should remove these RCTs from ones updated meta-analysis. Furthermore, the bias does, in our scenario, not stem from bias from the previous meta-analysis. The bias comes from the decision-process and this point can be illustrated by re-visiting Figure 6.1. The top plot in this figure shows an unbiased sample of test statistics under the null hypothesis of $\theta = 0$. None of these samples are biased, they are simply samples. However, once the decision to continue only those which points in favour of the intervention, we create a mechanism that will push the point estimate of the updated meta-analysis towards an untrue larger effect (bottom plot on Figure 6.1) even when the new evidence is not biased.

Manuscript III aims to provide an estimator which can adjust for this decision-process. The motivating example just described encapsulates when we would consider it appropriate to investigate adjusting ones point estimate. In the scenario where new trials are motivated by earlier evidence's result and direction, and are then combined with the earlier evidence, we suggest that the updated meta-analysis point estimate is investigated for potential bias due to the decision to continue based on promising results. With this description of the motivation we conclude the introduction to Manuscript III.

7. Conclusions

This chapter presents a summary of the main contributions and proposes some areas for future research which are a natural continuation of this thesis.

7.1 Main contributions

This section provides a brief summary of the three papers comprised in this thesis. Moreover it contains the main contributions of the papers. The papers are placed in Chapter 8.

In the first manuscript, we showed that linear mixed models provide a novel and useful model framework for subgroup analysis in meta-analysis, especially in the presence of incompletely reported data. Manuscript I gives a comprehensive study of current methods for subgroup meta-analysis, as well as the proposed linear mixed models, for estimation of the interaction effect, also called the effect modifier. Based on theoretical and simulation results provided in the paper, we provided recommendations for when to use which method as well as introducing a new model framework to fill the gaps between currently existing methods.

In the second manuscript, we implemented and updated Trial Sequential Analysis (TSA) in R. The original software is no longer updated to the latest gold standards and it is impossible to run simulations to validate the methods used in the software. A comprehensive implementation of TSA methods is presented for the R computing environment, in the package RTSA, which is a modification and expansion of the original software implemented in Java. The modification involved implementation of several new features such as sample size adjustments due to a sequential design, changing the one-sided designs with futility boundaries to consistent tests and implementing binding futility boundaries. Further to increase ease of use, several package vignettes were written and is an addition to the presented manuscript.

In the third manuscript, we investigated whether sequential conditional bias due to the decision to continue a meta-analysis could be adjusted for. Inspired by bias adjustment methods from group sequential trials, we developed a new estimator for sequential meta-analysis. The estimator conditions on the sample path of the test statistic, where it is expected that the sampling process was influenced by the sign of the point estimate from the previous analysis and that the analysis was insignificant. Based on the simulation results, recommendations were given as to when the proposed estimator is most useful.

7.2 Ongoing and future research

This section describes the future research stemming from this thesis.

Subgroup meta-analysis

After the publication of Manuscript I: “Linear Mixed Models for investigating effect modification in subgroup meta-analysis” (Soerensen et al., 2023a), another paper investigating effect modification in meta-analysis has been published (Godolphin et al., 2022). We note that this paper also address similar issues to those investigated in our paper. One of the main differences is the latter paper’s assumption of no ecological bias or aggregation bias. Presence of ecological bias is a prime motivating issue in Manuscript I.

Trial Sequential Analysis

A top priority for future research is to investigate full control of type I and type II errors. This includes comparing the method to other methods such as naive updating of the meta-analysis, semi-Bayesian and fully Bayesian methods. Here it is of interest to also look at the stability of the estimation of heterogeneity as the Bayesian methods uses a prior distribution to assist the modelling of the parameter. Some of the work has already started and is presented in Appendix B.

Furthermore, the software now includes random-effects models based on both DerSimonian-Laird and the Hartung-Knapp-Sidik-Jonkmans adjustment to the DerSimonian-Laird. A guide on when to use each method, and whether one or the other is the more appropriate to be the default in the R package is a topic of future research.

Debate about sequential meta-analysis methods

A topic which has not been formally addressed in this thesis is the debate about the role of Trial Sequential Analysis. TSA is a well-accepted methodology with strong proponents, but has also been subject to critical assessments. Two documents were released from the Cochrane group (found at <https://methods.cochrane.org/methods-cochrane/repeated-meta-analyses>) where the recommendation about using sequential meta-analysis methods such as TSA is: “The Expert Panel recommends against the use of sequential methods for updated meta-analyses in most circumstances within the Cochrane context. They should not be used for the main analyses, or to draw main conclusions.”. This statement is cited from https://methods.cochrane.org/sites/methods.cochrane.org/files/uploads/tsa_expert_panel_guidance_and_recommendation_final.pdf. There are several reasons for why this statement was not addressed in more detail in this thesis.

A reason for not addressing the criticism in this thesis is that parts of the future research plan will investigate the validity of TSA under various scenarios as described in the section above and below. While we do not question that the statement from Cochrane is based on reasoned considerations, none of this research has been formally published. We want to be critically appraise the use of TSA, but this requires that we have concrete results to point to. We expect

that we might be critical in some scenarios but not all. This is mentioned in several places in Chapter 4, Chapter 5, Manuscript II and Appendix B. At the same time, this thesis has also demonstrated the applicability of TSA and its promising role in updated meta-analyses.

Another reason for not including more detailed assessment of the Cochrane recommendations is that the statement is within the Cochrane context, which may not always be the context of interest. This context seems to be close in analogy to the Living Systematic Review context where binary decisions based on p-values are not recommended. This might be the right paradigm for many meta-analyses. However, we believe that the results of many meta-analyses are used in a decision-making context based on hypothesis testing. This might not always be the most appropriate approach, but it is still widely applicable. Using sequential meta-analysis methods such as TSA is one way to provide a more conservative perspective on decision-making by trying to account for the updated hypothesis testing in an updated meta-analysis. This does not mean that we assume that it will always provide valid results. But it may serve in many scenarios as a powerful sensitivity analysis or even main analysis. Recommendations of when, if and how to use TSA or sequential meta-analysis in general is part of future research.

We plan to formally address the Cochrane statements, which we most likely are going to agree with in some scenarios, in the future research. We further note that GRADE recommends TSA as a valid methodology in systematic reviews and meta-analyses (Duhailib et al., 2024) which provides further support for TSA being a useful tool in meta-analysis.

RTSA

For now the RTSA package has been created for the purpose of recreating Trial Sequential Analysis. This means that the software can be used for creating TSA to the data of the user. As seen from the additional results located in Appendix B, TSA cannot always guarantee the control of the type-I- and type-II-error. In this scenario it might still be of interest to use the method but with additional information about how likely it is to have a type-I-error and a type-II-error under varying assumptions. Dissemination of how to use the software for simulation is also part of future research.

Several methods for inference are included in the software. Some of these were presented in Chapter 6. A vignette discussing the different types of inference is part of planned additions to the package. This includes what the difference is between a TSA-adjusted confidence interval, the naive confidence interval and the stage-wise confidence interval.

Effect smaller than minimal clinical relevant value

There is a clear connection between the rejection of the null hypothesis and the values contained in a confidence interval. A similar connection might be anticipated between entering the futility area and having a confidence interval that excludes the minimal clinical value. This is however not the case. Additional boundaries or other type of boundaries than futility boundaries might be interest to include in TSA. These new boundaries should be distinct when a 95% confidence interval does not contain the minimal clinical relevant value.

7.3 Summary

In summary, this thesis has made novel contributions to theoretical and computational methodology for meta-analysis. Methods for subgroup and sequential analysis within meta-analyses have undergone recent evolution and we have identified a number of areas for innovation and further development. The research presented here has provided new ways to conduct subgroup analysis using a comprehensive analytical framework that incorporates and enhances existing approaches. It has also provided analytical and computational tools for addressing issues related to sequential sampling and inference in cumulative meta-analysis. As well the advances presented in the three original manuscripts, the research presented here will provide the basis for further developments in future research.

8. Manuscripts

8.1 Manuscript I

Linear mixed models for investigating effect modification in subgroup meta-analysis

Anne Lyngholm Soerensen & Ian C. Marschner

Details: Published in *Statistical Methods in Medical Research* in 2023.

Linear mixed models for investigating effect modification in subgroup meta-analysis

Statistical Methods in Medical Research
2023, Vol. 32(5) 994–1009
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09622802231163330
journals.sagepub.com/home/smm



Anne Lyngholm Sørensen^{1,2}  and Ian C Marschner³ 

Abstract

Subgroup meta-analysis can be used for comparing treatment effects between subgroups using information from multiple trials. If the effect of treatment is differential depending on subgroup, the results could enable personalization of the treatment. We propose using linear mixed models for estimating treatment effect modification in aggregate data meta-analysis. The linear mixed models capture existing subgroup meta-analysis methods while allowing for additional features such as flexibility in modeling heterogeneity, handling studies with missing subgroups and more. Reviews and simulation studies of the best suited models for estimating possible differential effect of treatment depending on subgroups have been studied mostly within individual participant data meta-analysis. While individual participant data meta-analysis in general is recommended over aggregate data meta-analysis, conducting an aggregate data subgroup meta-analysis could be valuable for exploring treatment effect modifiers before committing to an individual participant data subgroup meta-analysis. Additionally, using solely individual participant data for subgroup meta-analysis requires collecting sufficient individual participant data which may not always be possible. In this article, we compared existing methods with linear mixed models for aggregate data subgroup meta-analysis under a broad selection of scenarios using simulation and two case studies. Both the case studies and simulation studies presented here demonstrate the advantages of the linear mixed model approach in aggregate data subgroup meta-analysis.

Keywords

Study-level confounding, ecological bias, effect modification, linear mixed models, subgroup meta-analysis

1 Introduction

Subgroup meta-analysis can be used for comparing treatment effects between participant-level subgroups. Subgroups are defined to be different levels of a baseline characteristic. If differential effects are found based on subgroup, the subgroup variable is then a treatment effect modifier. Using this knowledge, treatment can be tailored to the different subgroups which is a way to personalize medicine. Another reason for conducting subgroup meta-analysis is to investigate sources of heterogeneity that might influence the meta-analysis on a higher level. This includes the overall pooled treatment effect estimate across subgroup levels. In this article, we will be concerned with detecting possible treatment effect modifiers (interaction effects) in meta-analysis, while also addressing heterogeneity estimation.

The interaction effect estimated using meta-analysis methods can be split into two types: across- and within-study interaction effects.¹ The across-study interaction effect represents the overall interaction effect as a function of subgroup

¹School of Mathematical and Physical Sciences, Macquarie University, Sydney, Australia

²Section of Biostatistics, Department of Public Health, University of Copenhagen, Denmark

³NHMRC Clinical Trials Centre, University of Sydney, Sydney, Australia

Corresponding author:

Anne Lyngholm Sørensen, Section of Biostatistics, Department of Public Health, Øster Farimagsgade 5 ogg. B, Postboks 2099 1014 København K, Denmark.

Email: als@sund.ku.dk

characteristics such as the subgroup fraction. It is a weighted average across all studies which can be used for general statements about the interaction effect on an ecological level. In comparison, the within-study interaction effect is useful for statements on an individual level. When study-level confounding, also called ecological bias, is present, the across- and within-study interaction effects will differ. The within-study interaction effect, and its use in studying individual-level effect modification, will be the focus of this article.

Estimation of within-study interaction has been mostly studied within individual patient data (IPD) meta-analysis using case studies or simulation studies. IPD subgroup meta-analysis was recommended over aggregate data (AD) subgroup meta-analysis.² For scenarios with a mix of IPD and AD, some authors have recommended not to use AD in the modeling of the within-study interactions.³ Considering a scenario where only very few studies are able to provide IPD, the decision not to use AD causes a potential great loss of information and power. If only AD data is available, meta-analysis of the interaction effects per study has been recommended over other methods.⁴ This method can be called interaction meta-analysis (IMA).

The recommendation to use IMA for AD subgroup meta-analysis is based on results from IPD subgroup meta-analysis studies and the property of the method that it has an unbiased estimate of the within-study interaction effect.⁴ Hence the method is robust under study-level confounding. Provided that we have access to studies which provide information on all subgroups of interest, IMA is an appropriate method to estimate the interaction effect. An obvious drawback of the method is that it can only include studies with all subgroups present. Trials which provide only a subset of the subgroups are excluded. At the other end of the spectrum we could have studies that would never provide both subgroups of interest. An example is the effect of a specific treatment where location of the study (e.g. the US vs. Europe) defines the subgroups. IMA cannot accommodate studies if none of the studies contains both subgroups. One would instead fit a subgroup specific meta-analysis, where a meta-analysis is done per subgroup and the pooled estimates are compared. This approach is recommended in the Cochrane Handbook.⁵ Subgroup specific meta-analysis is not robust towards study-level confounding and had any studies had both subgroups, this information and possible dependence would not be used in the model.

We will propose a new model to fill the gap between IMA and subgroup specific meta-analysis, using linear mixed models (LMMs). In general using a mixed effects model setup allows for a lot more flexibility compared to traditional meta-analysis methods. Investigating different sources of heterogeneity within the same model becomes easier. For subgroup meta-analysis there may be variation not only specific to the study but also specific to the subgroup. The models we suggest in this article will provide the possibility to estimate both the between-study variation and the within-subgroup variations. Another advantage of using LMMs is the possible ways to deal with pooled and missing data. By pooled data, we refer to situations where some of the studies appropriate for a subgroup meta-analysis might not provide information other than the overall treatment effect (pooled estimate for all subgroups) and the fraction of the sample within each subgroups of interest. Hence in these studies, we are not given subgroup specific estimates, yet they may still contain information that can make subgroup analysis more efficient. Furthermore, if we wish to adjust for other variables in the meta-analysis that might not be provided for all studies, we are in a missing data situation. This can be handled by LMMs and both scenarios will be studied in this article. Although LMMs have been used for meta-analysis,⁶ their use as a general methodology for subgroup meta-analysis has not previously been investigated. This article will do this focusing on interaction effects and how heterogeneity can be flexibly modeled when arising from different sources of variation in AD subgroup meta-analysis. If there is sufficient IPD data available, the recommendation is to do IPD subgroup meta-analysis. Riley et al.⁷ provide a thorough guide on how to perform IPD meta-analysis of treatment-covariate effects.

This article begins with a short introduction to the two motivating examples. We then give a brief review of the methods currently used in AD subgroup meta-analysis followed by a presentation of the proposed LMM structure which we call the basic LMM. Methods for fitting a LMM for meta-analysis are then presented. Next, we present two case studies based on the motivating examples analysed with the existing and proposed methods. Finally, the article concludes with a simulation study which compares the existing models with each other and the proposed model under varying scenarios.

2 Motivating examples

We present two motivating examples that will be used as case studies in this article. The first case study has been used in another methodological article concerning subgroup meta-analysis, and is concerned with early supported discharge (ESD) from hospital.⁴

ESD services aim to allow patients to return home from hospital earlier than usual and receive more rehabilitation in the familiar environment of their own home.⁸ An earlier discharge may accelerate return to home and provide better patient and carer outcomes such as improvements in mood, increased activities of daily living and subjective health status. A subgroup meta-analysis of whether the effect of ESD was differential depending on presence of carer versus non-presence of carer has been investigated in a Cochrane systematic review⁸ and in a methodological paper by Fisher et al.⁴ In this case study,

previous analyses used IMA excluding any studies presenting results only for one of the subgroups. Here we use LMMs to include all available information.

The second case study is concerned with the effectiveness of immune checkpoint inhibitors (ICIs) in cancer therapy. ICIs, also known as a type of immunotherapy, have substantially improved outcomes for many advanced cancers. However, only a subset of patients may respond to these therapies. For cancers such as non-small cell lung cancer (NSCLC), head and neck cancer (HNC), and urothelium cancer (UC), tobacco smoking is strongly implicated in tumour mutagenesis. For NSCLC and HNC, smoking is also associated with greater tumor mutation burden.⁹ Lee et al.⁹ used subgroup specific meta-analysis to test whether the effect of ICI is differential depending on smoking status. Their aim was to investigate whether smoking status is useful for quantifying the effect of ICI therapy. We will here also investigate the interaction effect using LMMs, which provide a more flexible and potentially more valid approach than a subgroup specific analysis.

3 Methods for subgroup meta-analysis

In this section, we begin with an overview of existing methods for AD subgroup meta-analysis before presenting the proposed LMM. The Cochrane Handbook⁵ mentions two methods for studying subgroups and investigating whether the treatment effect is differential depending on subgroup level. The methods are subgroup specific meta-analysis and meta-regression. IMA is not mentioned, but it has been recommended over the other two methods for estimating the within-study interaction effect.⁴ We begin with a brief overview of the three approaches before presenting the proposed LMM.

Subgroup specific meta-analysis starts with estimating the pooled treatment effect in each subgroup. That is, a standard meta-analysis is conducted within each subgroup, separately. The interaction effect is then calculated by taking the difference between the pooled subgroup specific treatment effects. Heterogeneity can be modeled for each subgroup by estimating the between-study variance within each subgroup meta-analysis. A visualization of the method is given in the forest plot in Figure 1. Figure 1 is based on data from the second motivating example. Fisher et al.⁴ refer to the subgroup specific meta-analysis method as the “deluded” method.

Subgroup meta-regression uses the pooled treatment effect estimate and a covariate for subgroup which can be categorical or numeric. Meta-regression using categorical subgroups regresses the pooled treatment effect against the subgroup proportion, for example, proportion of males. It is strongly recommended not to use this approach for estimating the within-study interaction effect in subgroup meta-analysis because of the potential ecological bias.⁴ For this reason, we will not be using this method further in this article. Fisher et al.⁴ refer to subgroup meta-regression as the “daft” method.

Interaction meta-analysis (IMA) calculates the interaction effect per study for the studies which provide the treatment effect for both subgroups of interest. The study interaction effects are then used as input in a standard meta-analysis. Heterogeneity can be modeled by adding a random study effect hence allowing for between-study variation. Fisher et al.⁴ refer to IMA as the “deft” method.

IMA has been recommended over subgroup specific meta-analysis as the interaction effect will be unbiased using this method.⁴ An early example of IMA can be found in Adelstein et al.¹⁰ In this article, we will compare our proposed LMM model to the IMA and the subgroup specific meta-analysis. There are scenarios where the subgroup specific meta-analysis is still valid, such as when no studies have more than one subgroup, comparisons to this method will also be made in the simulation studies. This scenario represents one end of the spectrum. At the other end if all studies provided all subgroups, the method of choice should be IMA. In the simulation study, we will consider both extremes and scenarios in between, including situations where we have a mix of studies where some report all subgroups while others only report single subgroups and studies only reporting the overall treatment effect. In the next subsection, we present the proposed model and then show how the two existing methods for AD subgroup meta-analysis can be incorporated into the overarching LMM.

3.1 Basic LMM

We start by specifying the basic form of our proposed LMM for estimating the within-study interaction effect. Let $\hat{\theta}_{kj}$ be the treatment effect for subgroup k ($k = 1, \dots, K$) in study j ($j = 1, \dots, J$), with observed standard error s_{kj} . Our proposed model is then

$$\hat{\theta}_{kj} = \mu + \alpha_k x_{kj} + B_j + G_{kj} + \varepsilon_{kj} \quad (1)$$

Here μ is the treatment effect of the reference subgroup, α_k is the difference in treatment effect (the interaction effect) compared with the reference subgroup with $\alpha_1 = 0$, x_{kj} is the indicator of subgroup k , B_j is a random study effect distributed with a $\mathcal{N}(0, \tau^2)$ distribution, G_{kj} is a random subgroup effect distributed as $\mathcal{N}(0, \tau_k^2)$ and ε_{kj} is the error term with a $\mathcal{N}(0, \sigma_{kj}^2)$ distribution. As we observe the estimated standard errors per trial s_{kj} , we replace σ_{kj}^2 with its estimate s_{kj}^2 . We therefore use $\hat{\sigma}_{kj}^2 = s_{kj}^2$ in place of the unknown σ_{kj}^2 which makes the model identifiable.

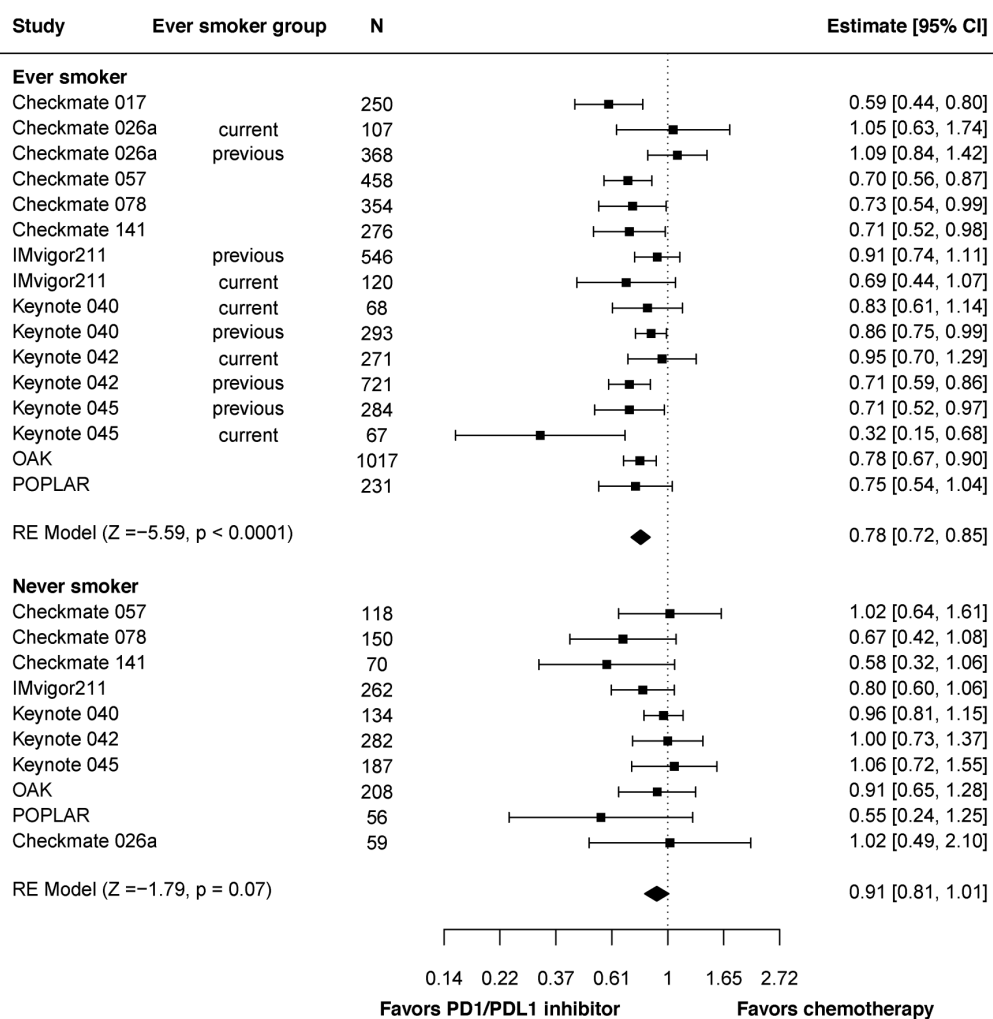


Figure 1. Forest plot of hazard ratios for overall survival comparing immune checkpoint inhibitors versus chemotherapy in ever-smokers and never-smokers participant subgroups. Hazard ratio for each trial is represented by the square and the horizontal line crossing the square represents the 95% confidence interval (CI). The diamond represents the pooled overall effect size estimated using a fixed-effect meta-analysis. All statistical tests were two-sided.

Assuming that the error terms ϵ_{kj} are independent, the random study effect B_j in model (1) induces positive correlation between treatment effects from the same study. One flexible feature of this basic LMM is that it can be easily modified to allow more general correlation structures. For example, by allowing error terms from the same study to be correlated, the basic LMM can allow for negative correlation of treatment effects from the same study. This flexible correlation structure is not available in any other previous methods proposed for AD subgroup meta-analysis. We will see why this is a useful modelling feature in one of the case studies.

Note that the basic model could resemble an arm-based network meta-analysis model of only direct comparisons. Arm-based network meta-analysis models the treatment arms of the different trials, where we here model the subgroup treatment effects. The focus in a network meta-analysis is to get consistent estimates of both direct and indirect comparisons between treatments. In a subgroup meta-analysis, we are interested in the contrast between the treatment effect in a reference subgroup and the other subgroups individually. This makes the network meta-analysis issue of consistency less applicable to subgroup meta-analysis. See Hong et al.¹¹ for a description of Bayesian arm-based network meta-analysis.

Notice also that we can have any number of categories of the subgroup variable. We will often be interested in two subgroups, but the method allows for multiple categories of a given patient characteristic.

3.2 Existing methods as LMMs

The correspondence between the existing subgroup meta-analysis methods and the basic LMM can be seen by expressing the existing methods in the LMM form (1).

The LMM version of subgroup specific meta-analysis is fitted in one-stage and can be defined as

$$\hat{\theta}_{kj} = \mu + \alpha_k x_{kj} + G_{kj} + \varepsilon_{kj} \quad (2)$$

As can be seen by comparing (1) with (2), subgroup specific meta-analysis is the same model as the basic LMM without the random study effect B_j . Having no study random effect B_j removes the dependence between subgroups from the same study and reduces (1) to (2). Fitting the model (2) allows for replication of the results from the traditional subgroup specific meta-analysis. Methods for fitting the model using standard statistical software is presented in the Supporting Information.

For IMA, we calculate the interaction effect between two levels of the subgroup within each study. Suppose we are interested in the difference in treatment effect between subgroups 1 and 2. Under the assumption that the subgroup treatment effect estimates $\hat{\theta}_{1j}$ and $\hat{\theta}_{2j}$ are normally distributed, we can let $\hat{\lambda}_j^{1,2} = \hat{\theta}_{2j} - \hat{\theta}_{1j}$, which is then the difference in treatment effect between subgroups 1 and 2 with subgroup 1 as a reference. The LMM version of the IMA can then be written as

$$\begin{aligned} \hat{\lambda}_j^{1,2} &= \hat{\theta}_{2j} - \hat{\theta}_{1j} \\ &= \alpha_2 + (G_{2j} - G_{1j}) + (\varepsilon_{2j} - \varepsilon_{1j}) \\ &= a + b_j + e_j \end{aligned} \quad (3)$$

Here $a = \alpha_2$ is the interaction effect, $b_j = G_{2j} - G_{1j} \sim \mathcal{N}(0, \tau_2^2 + \tau_1^2)$ is a study specific random effect of the interaction effects with $b_j \sim \mathcal{N}(0, \tau^2)$, and $e_j = \varepsilon_{2j} - \varepsilon_{1j} \sim \mathcal{N}(0, \sigma_{2j}^2 + \sigma_{1j}^2)$ with $e_j \sim \mathcal{N}(0, \sigma_j^2)$. As described in Section 3.1, given that we observe the standard errors s_{kj} , we set $\sigma_j^2 = s_j^2$. By (3) it can be shown that IMA is analogous to (1) with $\tau^2 = \tau_1^2 + \tau_2^2$ and $\sigma_j^2 = \sigma_{1j}^2 + \sigma_{2j}^2$.

We will see that using the basic LMM, we can fit scenarios where we have incomplete data. Such scenarios cannot be handled by the existing methods. Incomplete data and ecological bias can be handled using LMMs with extensions to the basic model.

3.3 Study-level confounding

Greenland and Morgenstern¹² define two sources of ecological bias. It can come from effect modification by the subgroup or from the subgroup acting like a confounder, a study-level confounding. As we are investigating the effect modification, we are worried by the latter type of bias. Firebaugh¹³ and Hua¹⁴ amongst others treats the confounding from the subgroup as a association between the rate or mean level of the subgroup and the endpoint. This will also be how we treat ecological bias. Ecological bias in subgroup meta-analysis will cause the estimated interaction effect to be biased in the basic LMM (1), when there is incomplete data, and in the subgroup specific meta-analysis under both complete and incomplete data. Study-level confounding is related to the study-specific fraction of the subgroup such as the fraction of males in the specific study. Hua et al.¹⁴ give a comprehensive discussion of study-level confounding. To ensure unbiasedness in the proposed LMM with two subgroups, an additional term β can be added to (1)

$$\hat{\theta}_{kj} = \mu + \alpha_k x_{kj} + \beta p_j + B_j + G_{kj} + \varepsilon_{kj} \quad (4)$$

where

$$p_j = \frac{\text{number of participants in subgroup } k = 2}{\text{number of participants}} \quad (5)$$

Here β is the change in treatment effect as a function of the study-specific subgroup fraction p_j of subgroup 2. By modeling the proportion of the subgroup, we combine the basic LMM (1) with a meta-regression model. If there is any dependence between the treatment effect and the ecological level of the subgroup, the β term removes this dependence from the interaction term α_k . By adjusting for the subgroup fractions in (4) we have extended the basic LMM (1) to account for the effects of ecological bias.

The model assumes that the ecological bias comes from the ecological level of one of the subgroups compared to the rest. If there is suspicion of more than one subgroups ecological level causes bias (can occur when comparing more than two subgroups), the model should be adapted to accommodate this.

3.4 Studies without all subgroups

Studies which only studied a subset of the subgroups of interest can be appropriate to add to the meta-analysis to add information. An advantage of the proposed LMM, unlike the IMA, is that such studies enter without having all subgroups of interest. In addition, if we wish to model systematic differences between studies with only a subset of the subgroups and studies providing all subgroups of interest, we can add a term to distinguish the two. Let m_j be a variable for missing subgroup(s) in study j , where $m_j = 1$ if one or more subgroups are not reported for study j . For studies which provide all subgroup specific treatment effects, set $m_j = 0$. We then write the model as

$$\hat{\theta}_{kj} = \mu + \alpha_k x_{kj} + \delta m_j + B_j + G_{kj} + \varepsilon_{kj} \quad (6)$$

where δ is the difference in treatment effect between studies with all subgroups available and studies with not all subgroups available. Note that the δm_j term can be adjusted to the desired scenario by changing the definition for m_j or allowing for multiple adjustment terms to describe the specific structures of missingness.

3.5 Studies without subgroup specific estimates

Studies that do not provide information on subgroup specific treatment effects can also be added to the general model. Suppose that the J studies are split into two sets, J^P and J^{SG} , where J^P is the set of studies with only pooled treatment effects and J^{SG} is the set of the studies with subgroup specific treatment effects. We then include the pooled treatment effect estimates into the general model using the following setup:

$$\begin{aligned} \hat{\theta}_{kj} &= \mu + \alpha_k x_{kj} + B_j + G_{kj} + \varepsilon_{kj} & \text{for } j \in J^{SG} \\ \hat{\theta}_j &= \mu + \sum_k \alpha_k p_{kj} + B_j + \sum_k G_{kj} p_{kj} + \varepsilon_j^* & \text{for } j \in J^P \end{aligned} \quad (7)$$

Here $\varepsilon_j^* = \sum_k \varepsilon_{kj} p_{kj}$, $j \notin J^P$ refers to the studies which are in J^{SG} and p_{kj} is the proportion of participants in subgroup k out of the entire study population. The linkage between the subgroup specific estimates and the pooled estimate is that $\hat{\theta}_j$ is a weighted sum of the $\hat{\theta}_{kj}$. The information from the pooled studies enter the model, by modeling $\hat{\theta}_j$ as a subgroup specific treatment effect. This means it will be modeled together with $\hat{\theta}_{kj}$.

A further complication is that the proportion of the subgroups of interest might not be known. This is a missing data scenario. We will show an example of handling missing data using the basic LMM in one of the case studies.

4 Applications

We will use the two motivating examples to illustrate the flexibility of the proposed LMM and its extensions. The first case study shows how to incorporate studies that do not provide information on all subgroups and that the basic LMM can be adjusted to handle a scenario of negative correlation between the subgroup treatment effects. The second case study will illustrate how the model can be used in scenarios with inclusion of studies where only pooled treatment effects are provided and where the subgroup proportions are unavailable.

Notice that the endpoints for the two applications are, respectively, continuous and time-to-event. For binary endpoints, it is possible to re-format the AD into IPD and use a one-step approach to fit our proposed model using generalized LMMs. See Hua et al.¹⁴ for ecological bias and the use of the one-step approach.

4.1 ESD from hospital

Fisher et al.⁴ presented a study in which the interest is in whether the effect of ESD from hospital compared to the conventional service was differential based on the presence of a carer.⁸ The primary endpoint was the number of days at hospital. The subgroup meta-analysis included nine studies where eight provided information on both subgroups (carer and no carer). The interaction effect was estimated to -6.47 (95% CI: -13.65 to 0.71 , p -value = 0.077) using IMA, which was the method recommended by Fisher et al.⁴ We can reproduce the results using a fixed effect model. While we expect an unbiased estimate of the interaction effect in the interaction term meta-analysis, we do not obtain an estimate of the actual treatment effect in either of the subgroups. Furthermore, one study is excluded from the meta-analysis as it only contains one of the subgroups of interest. Hence there has been an exclusion of data that might be informative. The removal of the study is based on what data the model is able to handle rather than based on clinical

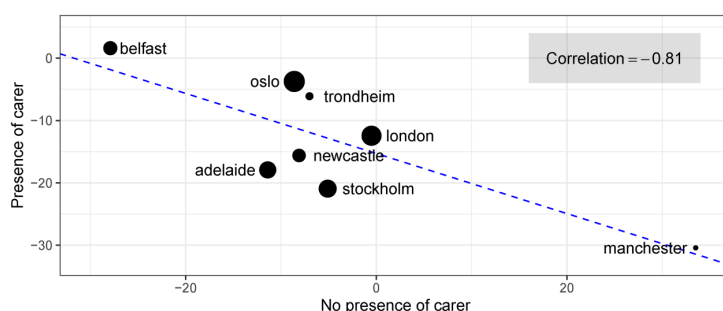


Figure 2. Scatterplot of relation between the effect of early supported discharge (ESD) for the carer and no carer subgroups. Only studies which provided information on both subgroups could be plotted. Points are sized according to the weight by the inverse of the variance. The blue dashed line is the regression line.

or quality considerations. We will investigate including the excluded study, which is called the Montreal study, in the basic LMM.

We will compare the original analysis of the case study to an analysis using the basic LMM (1) and some of its extensions. Before doing so, we note that the assumption of dependence between subgroups from the same study in (1) is modeled by a random effect of study. Modeling the dependence as a random effect forces the model to have a positive correlation for subgroups from the same study. This property may not always be appropriate as subgroups from the same study are not necessarily positively correlated and negative correlation can occur as we will see in this specific application. Fitting a block correlation structure instead of a study specific random effect is possible in LMMs, including the basic LMM (1). The correlation structure will allow subgroups from the same studies to be correlated but does not dictate a positive correlation. A possible explanation of negative correlation is the individual study centers might be expert in treating one of the subgroups but not the other. In the ESD study, the treatment effect in the carer subgroup is negatively correlated with the treatment effect in the no carer subgroup, which is shown in Figure 2 by plotting the effect estimates of the two subgroups against each other by study. Figure 2 indicates that assuming a positive correlation between subgroups from the same studies might be inappropriate. We will investigate what happens to the interaction effect by allowing for the extra study, Montreal, and changing from a study random effect to the block correlation structure. In principle, other correlation structures could also be fitted. See Supporting Information on how to fit the model in R.

Starting with fitting the basic LMM, where we include the Montreal study and model study as a random effect (dictating positive correlation), we get the model fit found in Table 1, Model 1. The standard deviation of the study random effect is estimated as < 0.0001 in the model. This could indicate independence between observations from the same study, but it could also indicate a non-positive correlation. If the correlation was truly negative, the study random effect variance would be pushed to 0. If we instead fit a block correlation structure, where we assume correlation between subgroups from the same study, we get the model fit found in Table 1, Model 2. We find in the basic LMM with the block correlation structure a correlation estimate of -0.81 and a decrease in AIC of 48.37 compared to 51.88. With this correlation structure, we perform a sensitivity analysis by including a term for the Montreal study, which is displayed in Table 1, Model 3. This model becomes the final model for this case study.

In the final model, we obtain a correlation estimate of -0.78 along with standard deviations of the subgroup random effects of 5.065 in the carer subgroup and < 0.0001 in the no carer subgroup. This information is useful for understanding where the heterogeneity exists in the data. Adding the Montreal study decreased the interaction effect compared to the analysis that excluded the Montreal study, see the estimated interaction effect of carer -3.29 in Model 1, Table 1 versus the Fisher estimate of -6.47 . But adding a parameter for the Montreal study, Model 3, Table 1, the interaction effect becomes more similar to the results found in Fisher et al.⁴ We additionally investigated potential study-level confounding/ecological bias by fitting (4), but found no evidence of this (p -value 0.2782). This contrasts with Fisher,⁴ but the ecological bias in that paper was concluded using fixed effect models. Changing the models from fixed effect to random effects generally decreases the influence of study-level confounding, as we shall see in the simulation study. A final comment about the basic LMM fit, is the possibility to also estimate the subgroup specific treatment effect which is found to be -6.82 (95% CI of -12.18 to -1.46) in the model for the no carer subgroup. This result comes from the final model 3.

Table 1. Results from three different versions of the basic LMM, all including the Montreal study. Model 1 is the basic LMM as formulated in (1). Model 2 is similar to Model 1 but fits a block correlation structure for the dependence between subgroups from the same study instead of fitting a random effect of study. Model 3 also fits the block correlation structure but compared to Model 2 adds a parameter for the Montreal study as formulated in (6).

	Coefficient	Estimate	95% CI	p-value
Model 1	Treatment effect for no carer	-6.11	(-11.66 to -0.55)	0.0354
	Interaction effect of carer	-3.29	(-12.23 to 5.65)	0.4133
Model 2	Treatment effect for no carer	-7.44	(-12.65 to -2.23)	0.0119
	Interaction effect of carer	-2.3	(-12.97 to 8.38)	0.6266
Model 3	Treatment effect for no carer	-6.82	(-12.18 to -1.46)	0.0197
	Interaction effect of carer	-5.05	(-16.37 to 6.26)	0.3261
	Montreal study	9.67	(-4.31 to 23.65)	0.1493

4.2 Effectiveness of ICIs

To investigate the efficacy of ICIs for lung cancer is modified by smoking status, Lee et al.⁹ conducted a meta-analysis of 11 studies which compared ICI therapy to chemotherapy based on smoking status. Figure 1 shows a replication of the forest plot. The study showed statistically significant evidence of an interaction effect between ICI therapy and smoking status, where ever smokers had a more beneficial treatment effect compared to never smokers. As the method used for estimating the interaction effect is subgroup specific meta-analysis, we will investigate whether similar conclusions are obtained using the basic LMM. Furthermore, as shown in Figure 1, the ever smoker category can be further split into two subgroups, current and previous smokers. Information on these subgroup levels was not available for all studies included in the original analysis. We will be using the basic LMM to investigate if there is an evidence for both of the ever smoker subgroups (current and previous) to have an increased treatment effect of ICI compared to the never smokers and whether the treatment effect between these two subgroups differs. As some of the ever smoker studies do not provide information on the current and previous smoker subgroups, we will use the basic LMM and the extension that handles studies without subgroup specific estimates (7). We will also investigate using missing data methods in this subsection in relation to the basic LMM.

4.2.1 Original analysis

We start by investigating the difference between the original subgroup specific meta-analysis and the proposed LMM analysis. Redoing the original analysis by Lee et al.,⁹ we obtain the results in Table 2, section Original analysis, Model 1.

There was no estimate of the interaction effect by Lee et al.,⁹ but the *p*-value was reported as 0.04, which is close to our estimate of 0.0339 and the conclusion remains the same. If we instead fit the basic LMM (1), we get the model fit in Table 2, Original analysis, Model 2. Here the conclusion is also the same as by Lee et al.,⁹ but with a borderline significant *p*-value. Hence including random effects on the subgroups gives similar conclusions to the original analysis. Notice, if we remove the random effects from the basic LMM, it would reduce to the fixed effects subgroup specific meta-analysis. We find that the results confirm the conclusions in the original analysis. In general, for a subgroup analysis such as this, the basic LMM is preferable to subgroup specific meta-analysis, however, in this case, it does not appear to have had any effect on the primary conclusions.

4.2.2 Previous and current smokers

We are also interested in splitting studies with the ever smoker category into previous and current smoker sub-categories. Six studies do not provide information on these categories. The studies missing the previous and current smoker subgroups are identifiable from Figure 1, as they are not already split into current and previous smokers. While none of these studies present the estimated treatment effects for these subgroups, the OAK¹⁵ and POPLAR¹⁶ studies provide the number of participants in the subgroups. Out of the 1017 ever smokers in the OAK study, there are 190 current smokers and the remaining 827 are previous smokers. In the POPLAR study, 46 participants are current smokers while 185 are previous smokers out of the total 231 ever smokers in the study.

We start our investigation of the finer subgroup categories with analysing just at the previous and current smokers subgroups and disregarding studies that do not provide the treatment effect estimates for the previous and current smokers. Fitting then the basic LMM we get results shown in Table 2, section Previous and current smokers, Model 1.

Including the OAK and POPLAR studies which provide the number of current and previous smokers in the study but not the treatment effects, can be handled by the basic LMM by the fitting model (7). The result is found in Table 2, section Previous and current smokers, Model 2. Neither of the models show a difference in the treatment effect between the previous

Table 2. Model fits from the “Effectiveness of immune checkpoint inhibitors” application.

	Coefficient	Estimate	95% CI	p-value
Original analysis				
Model 1	Treatment effect for never smokers	0.90	(0.81–1.01)	0.0743
	Interaction effect of ever smokers	0.88	(0.77–0.99)	0.0339
Model 2	Treatment effect for never smokers	0.91	(0.81–1.02)	0.0959
	Interaction effect of ever smokers	0.86	(0.74–1)	0.0504
Previous and current smokers				
Model 1	Treatment effect for previous smokers	0.84	(0.71–0.99)	0.0424
	Interaction effect of current smokers	0.98	(0.72–1.32)	0.8413
Model 2	Treatment effect for previous smokers	0.82	(0.72–0.93)	0.0118
	Interaction effect of current smokers	0.98	(0.73–1.32)	0.8758
Previous, current and never smokers				
Model 1	Treatment effect for never smokers	0.91	(0.8–1.02)	0.1018
	Interaction effect of current smokers	0.82	(0.61–1.09)	0.1533
	Interaction effect of previous smokers	0.87	(0.73–1.03)	0.0999
Model 2	Treatment effect for never smokers	0.91	(0.81–1.02)	0.0976
	Interaction effect of current smokers	0.81	(0.61–1.07)	0.1208
	Interaction effect of previous smokers	0.87	(0.75–1.02)	0.0798
Model 3	Treatment effect for never smokers	0.91	(0.81–1.01)	–
	Interaction effect of current smokers	0.82	(0.6–1.13)	–
	Interaction effect of previous smokers	0.87	(0.74–1.03)	–

and current smokers. This suggests that it is appropriate to combine the two subgroups as originally done by Lee et al.⁹ For illustrative reasons, we will to keep the two subgroups separate in the following subsection.

4.2.3 Previous, current, and never smokers

In this section, we will add the never smoker subgroup to the study. We start with including studies with complete reporting. Hence we will only consider those studies which provide the treatment effect estimate of all three subgroups. Fitting the basic LMM we get results shown in Table 2, section Previous, current, and never smokers, Model 1. We see from Table 2 that there is no statistically significant difference between the never smoker subgroup and, respectively, the current and previous smoker subgroups.

With an extension of the basic LMM, we can include the OAK and POPLAR studies using model (7). As we still do not have subgroup specific treatment for these two studies for the previous and never subgroups, we can only use the pooled treatment estimate. Fitting model (7), we get the results found in Table 2, section Previous, current, and never smokers, Model 2. Including the OAK and POPLAR studies using the basic LMM with extension (7) did not change the estimated effects.

Including the remaining studies, that did not split the ever smoker subgroup into current and previous smokers, can be done by using techniques to handle missing data. For example, using multiple imputations we can also use the studies which did not provide the number of current and previous smokers. To include these studies, we imputed the fraction of previous and current smokers based on the distribution of these subgroups within the ever smoker subgroup of the studies that reported all three subgroups. Setting the number of imputations to 20, we get the model fit found in Table 2, section Previous, current and never smokers, Model 3. Using a higher number of imputations did not change the overall results. We see that the estimates in the current and previous smoker groups moved further away from 1 indicating a more beneficial treatment effect for the current and previous smokers compared to the never smoker subgroup.

5 Simulation study

We conducted a range of simulation studies which will be reported in this section and in the Supporting Information. The simulations compare the existing methods with the basic LMM where our focus is on estimation of the interaction effect. Hence we consider the following models:

1. IMA (fixed and random effects).
2. Subgroup specific meta-analysis (fixed and random effects).
3. Basic LMM with and without extensions.

The basic LMMs are able to handle scenarios when reporting on subgroups is not complete and some, if not all, studies present results on more than one subgroup. This includes scenarios where subgroups are not reported for some studies and/or only the pooled estimate is reported. Such scenarios will be investigated in the simulation study, but we will also consider scenarios where we have complete reporting on all subgroups. As study-level confounding is a concern in subgroup meta-analysis, we will also perform simulations where we have study-level confounding. We will fit the basic LMM using the `nmle` package in R.¹⁷ Both the fixed effect and random effects versions of the IMA and the subgroup specific meta-analysis will be fit using the `metafor` package in R.¹⁸ A section on how to fit the basic LMM (1) with a fully reproducible example is provided in the Supporting Information. Although LMMs are a standard statistical model, there are a number of complexities that arise in their application to meta-analysis models. Using standard statistical software such as R and SAS,^{19,20} we need to coerce the software to fit a meta-analysis using LMMs.

5.1 Simulation models

We will simulate data on patient level, which we then summarize to represent the data on study level. To introduce the participant level, let $i = 1, \dots, n_{kj}$ be the indicator for participant i in subgroup k study j , where n_{kj} is the number of participants in subgroup k study j . Define $z_{ikj} \in \{0, 1\}$ to be the indicator of treatment, and $x_{ikj} \in \{0, 1\}$ to be the subgroup for participant i in study j subgroup k . We will be using two models for simulation. The first model includes subgroup effect modification without study-level confounding

$$\theta_{ikj} = \beta_{treatment} \cdot z_{ikj} + \beta_{subgroup} \cdot x_{ikj} + \beta_{treatment \times subgroup} \cdot z_{ikj} \cdot x_{ikj} + B_j + G_{kj} + \varepsilon_{ikj} \quad (8)$$

The second model includes both subgroup effect modification and study-level confounding

$$\begin{aligned} \theta_{ikj} = & \beta_{treatment} \cdot z_{ikj} + \beta_{subgroup} \cdot x_{ikj} + \beta_{treatment \times subgroup} \cdot z_{ikj} \cdot x_{ikj} \\ & + \beta_{study\text{-}confounding} \cdot z_{ikj} \cdot y_j + B_j + G_{jk} + \varepsilon_{ikj} \end{aligned} \quad (9)$$

Here $\beta_{treatment}$ is the treatment effect, $\beta_{subgroup}$ is the subgroup effect, $\beta_{study\text{-}confounding}$ is the study-level confounding and $\beta_{treatment \times subgroup}$ is the interaction effect, which we are interested in estimating. As in the basic LMM, B_j is the study random effect, where $B_j \sim \mathcal{N}(0, \tau^2)$ and G_{kj} is the subgroup random effect, where $G_{kj} \sim \mathcal{N}(0, \tau_k^2)$. As we are simulating on patient level the error is distributed as $\varepsilon_{ikj} \sim \mathcal{N}(0, \sigma_{kj}^2)$. Data will be simulated using model (8) and (9). When we investigate the fit of the models under study-level confounding, we will be simulating from model (9). Notice that model (8) is equal to model (9) when $\beta_{study\text{-}confounding} = 0$. Both models are related to models used in a paper by Hua et al.,¹⁴ where ecological bias was investigated in IPD meta-analyses models with a focus on survival data. We introduce ecological bias in the same way as Hua et al.,¹⁴ by the variable y_j . This variable is defined to be 1 when the reference subgroup fraction is more than 50% of the specific study j and 0 otherwise. This reflects one way of introducing bias into the model by assuming that the level of the subgroup is associated with an external effect on the endpoint.

We will investigate two subgroups, hence $k \in \{1, 2\}$, and a changing number of studies ranging from 6 to 20. For each scenario we simulated 10,000 meta-analyses. Table 3 shows which parameter values we will be considering. Besides the variation introduced by the two random effects and the error, we have additional variation in the model due to variation in subgroups sizes and study sizes. For more details see Supporting Information.

5.2 Simulation results

The simulation study considers three scenarios:

- Complete reporting
- Incomplete reporting: Missing subgroups
- Incomplete reporting: Pooled treatment effects.

When we are fitting the basic LMM we will use the study-level confounding extension in (4). For the incomplete reporting where not all studies report subgroup specific treatment effects, we will instead fit the basic LMM (7) but also include the study-level confounding adjustment. For a step-by-step guide and to see how the basic LMM performs without the study-level confounding adjustment, see Supporting Information.

Table 3. Parameter values used in the simulation study.

Parameters	Values
Number of studies	6, 10, or 20
Subgroups	$k = \{1, 2\}$
$\beta_{\text{treatment} \times \beta_{\text{subgroup}}}$	1
$\beta_{\text{treatment} \times \text{subgroup}}$	0.5
$\beta_{\text{study} - \text{confounding}}$	0, 0.5
B_j	$\mathcal{N}(0, 1.5)$
G_{kj}	$\mathcal{N}(0, 0.1)$ for $k = 1$, $\mathcal{N}(0, 1.5)$ for $k = 2$
ϵ_{ikj}	$\mathcal{N}(0, 1)$

5.2.1 Complete reporting

In this simulation study, all trials included report all subgroup specific treatment effects. We will report results for, respectively, six, 10, and 20 trials, $J \in \{6, 10, 20\}$. This range of trial numbers will also be used for subsequent simulation studies. The results of simulations with complete reporting can be found in Table 4. Notice that we present the mean estimate, the variation of the estimates, the bias (calculated as mean estimate minus the true value) and efficiency. Efficiency is calculated as the ratio of the mean squared error (MSE) between two models. The basic LMM's MSE will be set to the reference and set as the numerator. Hence efficiency below 100% means less efficient and efficiency above 100% means more efficient.

The simulations show that for completely reported subgroup data, the three methods of meta-analysis estimate the interaction effect equally well when there is no study-level confounding. When there is study-level confounding both the basic LMM and the IMA outperform the subgroup specific meta-analysis.

5.2.2 Incomplete reporting: Missing subgroups

This section contains the simulation results concerning scenarios where some studies do not report treatment effects for all subgroups. We consider missing subgroups in two ways:

1. 50% of subgroups missing in each simulated meta-analysis
2. Four subgroups missing in each simulated meta-analysis. For results, see Supporting Information.

The subgroups missing from each simulated meta-analysis are selected at random. Hence we can have scenarios where all removed subgroups are from one specific subgroup. To keep the number of studies fixed, we will not remove both subgroups from the same study.

For results on the first simulation scenario see Table 5, where it can be seen that the LMM approach performs at least as well as the other approaches. The only exception to this, is when there is no study-level confounding. In this case, the subgroup specific meta-analysis outperforms the other two methods. This has been noted earlier by Fisher et al.¹ and Hua et al.,¹⁴

5.2.3 Incomplete reporting: Pooled treatment effects

This section contains the simulation results concerning scenarios where some studies do not report subgroup specific treatment effects but provide the pooled estimate and the number of participants in each subgroup. We consider pooled treatment effects in two ways:

1. Pool 50% of subgroups per meta-analysis
2. Four studies are pooled in each simulated meta-analysis. For results, see Supporting Information.

For results on the first simulation scenario see Table 6. It can be seen that the basic LMM outperforms the other methods in all scenarios considered.

5.3 Summary of simulation results

We were able, by using simulation, to confirm the results of Fisher et al.⁴ in which the IMA was robust towards study-level confounding and the subgroup specific meta-analysis was sensitive towards study-level confounding. This is seen as the IMA model was not affected by the study-level confounding when estimating the interaction treatment effect. On the other hand, the subgroup specific meta-analysis estimation of the interaction effect was affected by study-level confounding.

Table 4. Simulating complete reported data with, respectively, six, 10, and 20 studies. Results are with and without study-level confounding. Efficiency (EFF) is calculated as the ratio between the mean squared error (MSE) of the basic LMM compared to the specific models. The basic LMM is used as the reference. FE stands for “fixed effect,” RE stands for “random effects,” and MA stands for “meta-analysis.”

	$\beta_{\text{study-confounding}} = 0$					$\beta_{\text{study-confounding}} = 0.5$				
	Mean	Var	Bias	EFF		Mean	Var	Bias	EFF	
J = 6										
FE subgroup specific MA	0.5003	0.0218	3×10^{-4}	103%		0.5613	0.0228	0.0613	85%	
RE subgroup specific MA	0.4998	0.0224	-2×10^{-4}	100%		0.5365	0.0242	0.0365	88%	
FE interaction MA	0.5001	0.0221	1×10^{-4}	101%		0.5001	0.0221	1×10^{-4}	102%	
RE interaction MA	0.4999	0.0227	-1×10^{-4}	99%		0.4999	0.0227	-1×10^{-4}	99%	
Basic LMM (4)	0.4999	0.0224	-1×10^{-4}	Ref.		0.4999	0.0225	-1×10^{-4}	Ref.	
J = 10										
FE subgroup specific MA	0.4997	0.0129	-3×10^{-4}	103%		0.5656	0.0134	0.0656	75%	
RE subgroup specific MA	0.4998	0.0131	-2×10^{-4}	101%		0.5415	0.0144	0.0415	83%	
FE interaction MA	0.4997	0.013	-3×10^{-4}	102%		0.4997	0.013	-3×10^{-4}	102%	
RE interaction MA	0.4997	0.0133	-3×10^{-4}	100%		0.4997	0.0133	-3×10^{-4}	100%	
Basic LMM (4)	0.4996	0.0132	-4×10^{-4}	Ref.		0.4995	0.0133	-5×10^{-4}	Ref.	
J = 20										
FE subgroup specific MA	0.4995	0.0062	-5×10^{-4}	103%		0.5692	0.0064	0.0692	57%	
RE subgroup specific MA	0.4996	0.0063	-4×10^{-4}	102%		0.5445	0.007	0.0445	72%	
FE interaction MA	0.4996	0.0063	-4×10^{-4}	101%		0.4996	0.0063	-4×10^{-4}	101%	
RE interaction MA	0.4997	0.0064	-3×10^{-4}	100%		0.4997	0.0064	-3×10^{-4}	100%	
Basic LMM (4)	0.4997	0.0064	-3×10^{-4}	Ref.		0.4996	0.0064	-4×10^{-4}	Ref.	

Table 5. Simulating incomplete reported data with, respectively, six, 10, and 20 studies. 50% of studies included are missing a subgroup. Results are with and without study-level confounding. Efficiency (EFF) is calculated as the ratio between the mean squared error (MSE) of the basic LMM compared to the specific models. The basic LMM is used as the reference. FE stands for “fixed effect,” RE stands for “random effects,” and MA stands for “meta-analysis.”

	$\beta_{study-confounding} = 0$					$\beta_{study-confounding} = 0.5$				
	Mean	Var	Bias	EFF	EFF	Mean	Var	Bias	EFF	EFF
J = 6										
FE subgroup specific MA	0.4997	0.0313	-3×10^{-4}	138%	138%	0.6903	0.0359	0.1903	65%	65%
RE subgroup specific MA	0.4996	0.0324	-4×10^{-4}	133%	133%	0.6757	0.0369	0.1757	69%	69%
FE interaction MA	0.5009	0.0461	9×10^{-4}	94%	94%	0.5009	0.0461	9×10^{-4}	101%	101%
RE interaction MA	0.5007	0.0481	7×10^{-4}	90%	90%	0.5007	0.0481	7×10^{-4}	97%	97%
Basic LMM (4)	0.4996	0.0432	-4×10^{-4}	Ref.	Ref.	0.4765	0.046	-0.0235	Ref.	Ref.
J = 10										
FE subgroup specific MA	0.4996	0.0172	-4×10^{-4}	141%	141%	0.7008	0.0192	0.2008	44%	44%
RE subgroup specific MA	0.4996	0.0177	-4×10^{-4}	138%	138%	0.6869	0.0198	0.1869	48%	48%
FE interaction MA	0.5	0.0265	0	92%	92%	0.5	0.0265	0	99%	99%
RE interaction MA	0.5004	0.0271	4×10^{-4}	90%	90%	0.5004	0.0271	4×10^{-4}	97%	97%
Basic LMM (4)	0.4991	0.0244	-9×10^{-4}	Ref.	Ref.	0.4708	0.0254	-0.0292	Ref.	Ref.
J = 20										
FE subgroup specific MA	0.4983	0.0088	-0.0017	132%	132%	0.7077	0.0096	0.2077	25%	25%
RE subgroup specific MA	0.4982	0.0089	-0.0018	131%	131%	0.6925	0.0099	0.1925	28%	28%
FE interaction MA	0.4983	0.013	-0.0017	89%	89%	0.4983	0.013	-0.0017	102%	102%
RE interaction MA	0.498	0.0132	-0.002	88%	88%	0.498	0.0132	-0.002	100%	100%
Basic LMM (4)	0.498	0.0116	-0.002	Ref.	Ref.	0.4675	0.0122	-0.0325	Ref.	Ref.

Table 6. Simulating incomplete reported data with, respectively, six, 10, and 20 studies. 50% of the studies included do not report subgroup specific treatment effects. Results are with and without study-level confounding. Efficiency (EFF) is calculated as the ratio between the mean squared error (MSE) of the basic LMM compared to the specific models. The basic LMM is used as the reference. FE stands for “fixed effect,” RE stands for “random effects,” and MA stands for “meta-analysis.”

	$\beta_{\text{study-confounding}} = 0$					$\beta_{\text{study-confounding}} = 0.5$				
	Mean	Var	Bias	EFF		Mean	Var	Bias	EFF	
J = 6										
FE subgroup specific MA	0.4994	0.0459	-6×10^{-4}	90%		0.5465	0.0478	0.0465	90%	
RE subgroup specific MA	0.4984	0.0481	-0.0016	86%		0.5227	0.051	0.0227	87%	
FE interaction MA	0.4992	0.0464	-8×10^{-4}	89%		0.4992	0.0464	-8×10^{-4}	97%	
RE interaction MA	0.4984	0.0486	-0.0016	85%		0.4984	0.0486	-0.0016	92%	
Basic LMM (4)	0.4901	0.0414	-0.0099	Ref.		0.4905	0.0449	-0.0095	Ref.	
J = 10										
FE subgroup specific MA	0.499	0.0269	-0.001	88%		0.5576	0.028	0.0576	83%	
RE subgroup specific MA	0.4989	0.0278	-0.0011	85%		0.5342	0.0299	0.0342	83%	
FE interaction MA	0.499	0.0272	-0.001	87%		0.499	0.0272	-0.001	95%	
RE interaction MA	0.499	0.0281	-0.001	84%		0.499	0.0281	-0.001	92%	
Basic LMM (4)	0.4893	0.0236	-0.0107	Ref.		0.49	0.0257	-0.01	Ref.	
J = 20										
FE subgroup specific MA	0.4992	0.0127	-8×10^{-4}	84%		0.5649	0.0132	0.0649	71%	
RE subgroup specific MA	0.4991	0.0129	-9×10^{-4}	83%		0.5409	0.014	0.0409	78%	
FE interaction MA	0.4995	0.0129	-5×10^{-4}	83%		0.4995	0.0129	-5×10^{-4}	95%	
RE interaction MA	0.4995	0.0131	-5×10^{-4}	81%		0.4995	0.0131	-5×10^{-4}	93%	
Basic LMM (4)	0.4787	0.0102	-0.0213	Ref.		0.4906	0.0122	-0.0094	Ref.	

Under complete reporting, we find that the IMA performs the best in terms of unbiasedness and variation of the interaction estimate. Table 4 shows that the basic LMM outperforms the subgroup specific meta-analysis when there is study-level confounding in terms of efficiency. As the basic LMM with study-level confounding adjustment performs similarly to the IMA model, the basic LMM can be used if we are interested in the treatment effect estimate in the reference subgroup, or are interested in investigating the more flexible heterogeneity structures that are also possible using the basic LMM (Table 4).

Under incomplete reporting, when some studies are without all subgroups, the subgroup specific meta-analysis performs well in terms of variance of the estimated interaction effect when there is no study-level confounding. Under study-level confounding, we find that the basic LMM performs better than the subgroup specific meta-analysis in terms of efficiency as the subgroup specific meta-analysis has efficiency of 25% to 69% of the basic LMM. When we let the number of studies missing a subgroup be half the total number of studies, we see as the total number of studies increases that the basic LMM consistently performs better than IMA (Table 5) when there is no study-level confounding. The efficiency of the interaction model is ~90% of the basic LMM. When the number of studies without all subgroups is fixed, we see that as the total number of studies increases, the IMA and the basic LMM starts to perform equally well, see Supporting Information.

Lastly we find under incomplete reporting, where some studies are not reporting subgroup specific treatment effects, that the basic LMM outperforms both the IMA and the subgroup specific meta-analysis in terms of the efficiency. This is seen for scenarios with and without study-level confounding. The efficiency of the interaction model is between 81% and 89% compared to the basic LMM under no study-level confounding and 92% and 97% under study-level confounding. When we increase the number of studies, but let the number of studies with pooled treatment effect be half the total number of studies, then increasing the total number of studies will increase the performance of the basic LMM compared to the other methods (Table 6). When the number of studies that do not report subgroup specific treatment effects are fixed and we increase the number of studies, the methods start to perform more equally when there is no study-level confounding, see Supporting Information.

We are able to make recommendations about when to use which model based on the results from the simulation studies. In the scenario of completely reported studies, one can choose between the IMA or the basic LMM with bias adjustment. For incompletely reported studies, where we include studies without all subgroups the subgroup specific meta-analysis performs best, when we are sure of no study-level confounding. In the more likely situation that we are uncertain about study-level confounding, the basic LMM with bias adjustment is the better choice. Finally, for the scenario with incompletely reported studies, where some studies only provide the pooled treatment effects, we recommend using the basic LMM with bias adjustment.

6 Discussion

Subgroup analysis in meta-analysis is recommended both for enabling a personalization of treatment to the specific groups and also for understanding the sources of heterogeneity. While it is generally recommended to use IPD over AD for conducting subgroup meta-analysis, the process of gathering enough IPD data to perform a meta-analysis can be resource intensive. An additional hurdle arises if not all trials are willing to share their IPD. Performing a subgroup meta-analysis using AD can be useful for investigating interaction effects prior to committing to an IPD subgroup meta-analysis. A guide on how to perform IPD subgroup meta-analysis can be found by Riley et al.⁷ While there has been recommendations for subgroup meta-analysis using AD, there is a lack of simulation studies to show how the different methods perform under varying scenarios. For the two existing methods presented in this article, the IMA and the subgroup specific meta-analysis, the first is appropriate when all studies report all subgroups. On the other hand, the latter is appropriate when no studies have more than one subgroup. For scenarios, in between we propose to use the basic LMM that unifies and extends these two approaches.

Using the basic LMM, we can incorporate incompletely reported studies which did not provide all subgroups of interest. These missing subgroups must be omitted from the IMA model. We can also incorporate studies that only provide the pooled treatment effect, which could not be incorporated in either the IMA or the subgroup specific meta-analysis. A further advantage is that the treatment effect in the reference subgroup is reported when using the LMM approach. Compared to the subgroup specific meta-analysis, the basic LMM can model dependence between subgroups from the same study and include methods for adjusting for study-level confounding. These advantages were visible in the case studies and simulation studies. In general using the LMMs for subgroup meta-analysis enables more flexibility to do sensitivity analyses when incorporating studies without all subgroup specific treatment effects and investigating study-level confounding.

The results from the simulation study generally provide support for using the basic LMM with bias adjustment for both complete or incomplete data. We find that the performance of the model is good when keeping the bias adjustment term, especially when we have incompletely reported data. In some cases, we may have some studies with IPD available and other studies with only AD. Future research will investigate how the basic LMM performs when extended to allow both IPD and AD in the model.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

ORCID iDs

Anne Lyngholm Sørensen  <https://orcid.org/0000-0002-8265-0394>

Ian C Marschner  <https://orcid.org/0000-0002-6225-1572>

Supplemental material

Supplemental material for this article is available online. Additional supporting information may be found online in the Supporting Information section of this article.

References

1. Fisher DJ, Copas AJ, Tierney JF, et al. A critical review of methods for the assessment of patient-level interactions in individual participant data meta-analysis of randomized trials, and guidance for practitioners. *J Clin Epidemiol* 2011; **64**: 949–967.
2. Belias M, Rovers MM, Reitsma JB, et al. Statistical approaches to identify subgroups in meta-analysis of individual participant data: a simulation study. *BMC Med Res Methodol* 2019; **19**: 183.
3. Riley RD, Steyerberg EW. Meta-analysis of a binary outcome using individual participant data and aggregate data. *Res Synth Methods* 2010; **1**: 2–19.
4. Fisher DJ, Carpenter JR, Morris TP, et al. Meta-analytical methods to identify who benefits most from treatments: daft, deluded, or deft approach? *BMJ* 2017; **356**: j573.
5. Higgins JPT, Deeks JJ and Altman DG. Analysing data and undertaking meta-analyses. In *Cochrane Handbook for Systematic Reviews of Interventions version 6.0*, chapter 10. 2019.
6. Cheung MWL. A model for integrating fixed-, random-, and mixed-effects meta-analyses into structural equation modeling. *Psychol Methods* 2008; **13**: 182–202.
7. Riley RD, Debray TPA, Fisher D et al. Individual participant data meta-analysis to examine interactions between treatment effect and participant? level covariates: statistical recommendations for conduct and planning. *Stat Med* 2020; **39**: 2115–2137.
8. Fearon P, Langhorne P. Services for reducing duration of hospital care for acute stroke patients. page nil, 2012.
9. Lee KWC, Lord SJ, Kasherman L, et al. The impact of smoking on the effectiveness of immune checkpoint inhibitors – a systematic review and meta-analysis. *Acta Oncol (Madr)* 2019; **59**: 96–100.
10. Adelstein BA, Dobbins TA, Harris CA, et al. A systematic review and meta-analysis of KRAS status as the determinant of response to anti-egfr antibodies and the impact of partner chemotherapy in metastatic colorectal cancer. *Eur J Cancer* 2011; **47**: 1343–1354.
11. Hong H, Chu H, Zhang J, et al. A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons. *Res Synth Methods* 2016; **7**: 6–22.
12. Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. *Int J Epidemiol* 1989; **18**: 269–274.
13. Firebaugh G. A rule for inferring individual-level relationships from aggregate data. *Am Sociol Rev* 1978; **4**: 557–572.
14. Hua H, Burke DL, Crowther MJ, et al. One-stage individual participant data meta-analysis models: estimation of treatment-covariate interactions must avoid ecological bias by separating out within-trial and across-trial information. *Stat Med* 2016; **36**: 772–789.
15. Fehrenbacher L, Pawel J, Park K, et al. Updated efficacy analysis including secondary population results for OAK: a randomized phase iii study of atezolizumab versus docetaxel in patients with previously treated advanced non-small cell lung cancer. *J Thorac Oncol* 2018; **13**: 1156–1170.
16. Fehrenbacher L, Spira A, Ballinger M, et al. Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised controlled trial. *The Lancet* 2016; **387**: 1837–1846.
17. Pinheiro J, Bates D, DebRoy S, et al. *nlme: linear and nonlinear mixed effects models*, 2020. R package version 3.1-151.
18. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw* 2010; **36**: 1–48.
19. Core Team R. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2020.
20. SAS Institute Inc. *SAS Software, Version 9.4*. Cary, NC. 2003 <http://www.sas.com/>.

Technical details I

The below supporting material in Manuscript I works as technical details.

Supporting information for "Linear mixed models for
investigating effect modification in subgroup meta-analysis" by
Anne L. Sørensen and Ian C. Marschner

1 Fitting the basic model using R or SAS

We will in this supporting information show how to fit the basic LMM model (1) in SAS and R. The example data is from the first application in this paper, "Early supported discharge (ESD) from hospital" also found in Fisher et al [5]. The data is also provided as supplementary material. We need to transform the data prior fitting the basic model, we will show how to do this in R.

1.1 R

The only necessary library to fit the basic LMM model in R is the nlme package [14]. We are going to assume that the data is ready for meta-analysis. We have the following variables in the data set, we name `d_example`:

Variable	Description
<code>study</code>	study
<code>te</code>	treatment effect
<code>te_se</code>	treatment effect's belonging standard error
<code>carer</code>	equal to 0 if non-carer subgroup, equal to 1 if carer subgroup
<code>no_carer</code>	equal to 0 if carer subgroup, equal to 1 if non-carer subgroup

We can then fit the basic model using the following code:

```
# read the nlme library
library(nlme)

# transform the variables for LMM fit
d_example$t_intercept = 1 / d_example$te_se
d_example$t_te = d_example$t_intercept * d_example$te
d_example$t_carer = d_example$t_intercept * d_example$carer
d_example$t_no_carer = d_example$t_intercept * d_example$no_carer

# fit the model - notice we constrain the residual error to 1
lme(t_te ~ t_intercept + t_carer - 1,          # model formula
    random = list(~ 0 + t_carer |study,       # carer specific random effect
                  ~ 0 + t_no_carer |study,    # non carer specific random effect
                  ~ t_intercept - 1 | study), # study random effect
    data = d_example, control = list(sigma = 1)) # constrain the residual error
```

1.1.1 Basic model with correlation pattern

We can fit a model with a correlation pattern using the following code:

```
lme(t_te ~ t_intercept + t_carer - 1,           # model formula
    random = list(~ 0 + t_carer |study,         # carer specific random effect
                  ~ 0 + t_no_carer |study),     # non carer specific random effect,
    correlation = corCompSymm(form = ~ t_intercept - 1 | study), # correlation pattern
    data = d_example, control = list(sigma = 1)) # constrain the residual error
```

There are multiple different correlation patterns in the `nlme` package in R [14], the above code is an example of one.

1.2 SAS

We wish to show two different ways to fit the subgroup meta-analysis. The first is a way to fit the basic LMM model (1) without the subgroup specific random effects and the second version shows how to fit the basic LMM model (1) with the subgroup specific random effects.

1.2.1 Basic model with no subgroup specific random effect

Where it is not possible to constrain the error variances in R, it is possible to constrain them in SAS. Let `d_var` be a data set containing a single variable, where the first observation is an arbitrary value of the variance of the random effect of study and the remaining observations are the actual estimated variances of the subgroup specific estimated treatment effects. We can then fit a model that constrain the error variances.

```
proc mixed data=d_example;
class study carer(ref = '0') _ ;
model te = carer;
random study;
REPEATED / GROUP = _;
parms / parmsdata = d_var EQCONS=2 to 18;
run;
```

Here `parms / parmsdata = d_var EQCONS=2 to 18;` sets the residuals equal to the observed variances.

1.2.2 Basic model with subgroup specific random effect

Since we can not fit the subgroup specific random effects using this model, we need to use the same trick as done in R. Assume that we have already transformed the variables as done in the R example. The SAS code for fitting the basic model is then:

```
proc mixed data=d_example;
class carer study;
model t_te = t_intercept t_carer / noint;
random t_intercept t_carer t_no_carer / sub = study;
parms (0) (0) (0) (1) / hold = 4;
run;
```

Here we constrain the residual error to 1 by using `parms (0) (0) (0) (1) / hold = 4;`.

2 Simulation step-by-step

2.1 Simulating studies where all subgroups are presented

2.1.1 Step 1- Set parameters

Define number of studies J , and set the remaining parameters from the parameter table provide in the paper. The following steps are then repeated 10000 times.

2.1.2 Step 2 - Generate IPD

From a uniform distribution sample the number of participants in the treatment groups $n_{j1,treatment} = n_{j2,treatment} \sim \text{Unif}(20, 120)$. We assume that there is equal number of participants per treatment group in the same subgroup per study, hence $n_{j1,treatment} = n_{j1,control}$. To introduce variation in the subgroup sizes, the proportion of subgroup 2 in relation to subgroup 1 can vary from -60% to +60%. Hence $n_{j2,treatment} = n_{j2,control} = n_{j1,treatment} \cdot \text{Unif}(0.4, 1.6)$. We then have $n_{j1,treatment} + n_{j1,control} + n_{j2,treatment} + n_{j2,control} = n_j$ which is the total number of participants for study j . Of these $n_{j1,treatment} + n_{j2,treatment}$ participants will have treatment $z_{ijk} = 1$, where the remaining are in the non-treatment group and will have $z_{ijk} = 0$. Further will $n_{j2,treatment} + n_{j2,control}$ participants be in subgroup 2 $x_{ijk} = 1$ and for the remaining participants (those in subgroup 1), $x_{ijk} = 0$. Finally we simulate j samples from normal distribution $B_j \sim \mathcal{N}(0, \tau^2)$ repeating each n_j times, and respectively sampling n_{j2} times from $G_{j2} \sim \mathcal{N}(0, \tau_2^2)$ and n_{j1} times from $G_{j1} \sim \mathcal{N}(0, \tau_1^2)$ per study j . Use then (7) or (8) for simulating θ_{ijk} depending on whether we are simulating study confounding.

2.1.3 Step 3 - Fit IPD and aggregate to AD

For each study, we run a linear regression model to estimate the treatment effects, $\hat{\theta}_{j1}$ and $\hat{\theta}_{j2}$, and their standard errors s_{j1} and s_{j2} . This data, the estimated treatment effects and their standard errors, will be used for the basic model and the two-stage subgroup meta-analysis. For the interaction meta-analysis, we need the difference in treatment effect. As our outcome is numeric and normally distributed, we simply calculate the difference in treatment effect $\hat{\lambda}_j$ and its standard error s_j as:

$$\hat{\lambda}_j = \hat{\theta}_{j2} - \hat{\theta}_{j1}, \quad \text{and} \quad s_j^2 = s_{j2}^2 + s_{j1}^2.$$

2.1.4 Step 4 - Fit the models

The last step prepares the data and fits the models. The R package metafor will be used for fitting respectively the fixed and random effects versions of the interaction meta-analysis and the fixed and random effects version of the subgroup specific meta-analysis. As the subgroup specific meta-analysis is a difference, the difference and belonging standard error is calculated.

For fitting the basic LMM model, we use R.

2.2 Simulating studies where not all subgroups are reported

When simulating studies where not all subgroups are reported, we start with deciding how many subgroups should be removed. This is either a fixed number (four) or a percentage (50%) of the studies included in the meta-analysis. The procedure is first to randomly pick which studies will only report on one subgroup, the next is then to randomly pick which of the subgroups are then removed. This is done for each simulation.

The same steps are then done as in subsection 2.1, with the exception of step 3 where we remove the selected subgroups. This means that we reduce the data for the basic LMM model and the subgroup specific meta-analysis where the number of removed subgroups are equal to the number of removed observations for these two models. For the interaction meta-analysis model, we will need to remove the studies missing one subgroup completely as it is not possible to include studies for which a difference can not be calculated.

2.3 Simulating studies without subgroup specific treatment effects

For simulating studies where, for some studies, we only have pooled treatment effect, we start with deciding how many studies should have pooled estimates. This is either a fixed number (four) or a percentage (50%) of the studies included in the meta-analysis. After selecting the studies, the pooled estimate and belonging standard error are calculated using the IPD per selected study.

The same steps are then done as in subsection 2.1, but with the expectation of step 3 where we do not fit the pooled data to the subgroup specific meta-analysis model and the interaction meta-analysis. Hence the data will be reduced for these two models. The basic model with extension to pooled data is used to fit the basic model (6). We also use the study-level confounding adjustment.

3 Simulation results

This section contains the additional simulations from section 6.3.2 and 6.3.3.

3.1 Incomplete reporting: Studies without all subgroups

This subsection contains the additional simulations from section 6.3.2.

$J = 6$	$\beta_{study-confounding} = 0$				$\beta_{study-confounding} = 0.5$			
	Mean	Var	Bias	EFF	Mean	Var	Bias	EFF
FE Two-stage SMA	0.5027	0.0353	0.0027	182%	0.7615	0.0425	0.2615	63%
RE Two-stage SMA	0.5027	0.0367	0.0027	175%	0.7515	0.0428	0.2515	66%
FE Interaction MA	0.5041	0.0714	0.0041	90%	0.5041	0.0714	0.0041	98%
RE Interaction MA	0.5038	0.0752	0.0038	85%	0.5038	0.0752	0.0038	93%
Basic LMM model (4)	0.5038	0.064	0.0038	Ref.	0.4861	0.0698	-0.0139	Ref.
$J = 10$	$\beta_{study-confounding} = 0$				$\beta_{study-confounding} = 0.5$			
	Mean	Var	Bias	EFF	Mean	Var	Bias	EFF
FE Two-stage SMA	0.498	0.0163	-0.002	123%	0.6654	0.0179	0.1654	48%
RE Two-stage SMA	0.4982	0.0167	-0.0018	120%	0.6486	0.0185	0.1486	54%
FE Interaction MA	0.4978	0.0215	-0.0022	93%	0.4978	0.0215	-0.0022	101%
RE Interaction MA	0.4981	0.0221	-0.0019	91%	0.4981	0.0221	-0.0019	99%
Basic LMM model (4)	0.4978	0.02	-0.0022	Ref.	0.4976	0.0217	-0.0024	Ref.
$J = 20$	$\beta_{study-confounding} = 0$				$\beta_{study-confounding} = 0.5$			
	Mean	Var	Bias	EFF	Mean	Var	Bias	EFF
FE Two-stage SMA	0.4999	0.007	-1e-04	110%	0.6158	0.0074	0.1158	40%
RE Two-stage SMA	0.4998	0.0071	-2e-04	109%	0.5943	0.0078	0.0943	49%
FE Interaction MA	0.5	0.008	0	97%	0.5	0.008	0	103%
RE Interaction MA	0.5	0.0081	0	96%	0.5	0.0081	0	102%
Basic LMM model (4)	0.4998	0.0077	-2e-04	Ref.	0.4805	0.0079	-0.0195	Ref.

Table 1: Simulating incomplete reported data with respectively six, ten and 20 studies all missing 4 subgroups. Results are with and without confounding. Efficiency (EFF) is calculated as the ratio between the mean squared error (MSE) of the basic LMM model compared to the specific models. The basic LMM model is used as the reference.

3.2 Incomplete reporting: Studies not reporting subgroup specific treatment effects

This subsection contains the additional simulations from section 6.3.3.

$J = 6$	$\beta_{study-confounding} = 0$				$\beta_{study-confounding} = 0.5$			
	Mean	Var	Bias	EFF	Mean	Var	Bias	EFF
FE Two-stage SMA	0.5021	0.0727	0.0021	86%	0.5361	0.0748	0.0361	91%
RE Two-stage SMA	0.5012	0.0775	0.0012	80%	0.5142	0.0806	0.0142	86%
FE Interaction MA	0.5019	0.0731	0.0019	85%	0.5019	0.0731	0.0019	95%
RE Interaction MA	0.5009	0.0782	9e-04	80%	0.5009	0.0782	9e-04	89%
Basic LMM model (4)	0.4914	0.0623	-0.0086	Ref.	0.4928	0.0693	-0.0072	Ref.
$J = 10$	$\beta_{study-confounding} = 0$				$\beta_{study-confounding} = 0.5$			
	Mean	Var	Bias	EFF	Mean	Var	Bias	EFF
FE Two-stage SMA	0.4992	0.0222	-8e-04	91%	0.5605	0.023	0.0605	81%
RE Two-stage SMA	0.4991	0.0228	-9e-04	89%	0.5366	0.0247	0.0366	83%
FE Interaction MA	0.4991	0.0225	-9e-04	90%	0.4991	0.0225	-9e-04	96%
RE Interaction MA	0.499	0.0232	-0.001	87%	0.499	0.0232	-0.001	93%
Basic LMM model (4)	0.4904	0.0201	-0.0096	Ref.	0.4909	0.0215	-0.0091	Ref.
$J = 20$	$\beta_{study-confounding} = 0$				$\beta_{study-confounding} = 0.5$			
	Mean	Var	Bias	EFF	Mean	Var	Bias	EFF
FE Two-stage SMA	0.5005	0.0078	5e-04	96%	0.5693	0.0081	0.0693	60%
RE Two-stage SMA	0.5005	0.0079	5e-04	95%	0.5447	0.0087	0.0447	72%
FE Interaction MA	0.5006	0.008	6e-04	95%	0.5006	0.008	6e-04	97%
RE Interaction MA	0.5007	0.0081	7e-04	93%	0.5007	0.0081	7e-04	96%
Basic LMM model (4)	0.4955	0.0075	-0.0045	Ref.	0.4966	0.0077	-0.0034	Ref.

Table 2: Simulating incomplete reported data with respectively six, ten and 20 studies. Four studies does not report subgroup specific treatment effects. Results are with and without confounding. Efficiency (EFF) is calculated as the ratio between the mean squared error (MSE) of the basic LMM model compared to the specific models. The basic LMM model is used as the reference.

4 Additional simulation results

In this section, we do not add the study-level confounding adjustment term to any of the models.

4.1 Complete reporting

$J = 6$	$\beta_{study-confounding} = 0$				$\beta_{study-confounding} = 0.5$			
	Mean	Var	Bias	EFF	Mean	Var	Bias	EFF
FE Two-stage SMA	0.5003	0.0218	3e-04	102%	0.5613	0.0228	0.0613	90%
RE Two-stage SMA	0.4998	0.0224	-2e-04	99%	0.5365	0.0242	0.0365	93%
FE Interaction MA	0.5001	0.0221	1e-04	100%	0.5001	0.0221	1e-04	108%
RE Interaction MA	0.4999	0.0227	-1e-04	98%	0.4999	0.0227	-1e-04	105%
Basic LMM model (1)	0.4998	0.0222	-2e-04	Ref.	0.5255	0.0232	0.0255	Ref.
$J = 10$	$\beta_{study-confounding} = 0$				$\beta_{study-confounding} = 0.5$			
	Mean	Var	Bias	EFF	Mean	Var	Bias	EFF
FE Two-stage SMA	0.4997	0.0129	-3e-04	101%	0.5656	0.0134	0.0656	81%
RE Two-stage SMA	0.4998	0.0131	-2e-04	99%	0.5415	0.0144	0.0415	89%
FE Interaction MA	0.4997	0.013	-3e-04	100%	0.4997	0.013	-3e-04	110%
RE Interaction MA	0.4997	0.0133	-3e-04	98%	0.4997	0.0133	-3e-04	107%
Basic LMM model (1)	0.4998	0.013	-2e-04	Ref.	0.5247	0.0137	0.0247	Ref.
$J = 20$	$\beta_{study-confounding} = 0$				$\beta_{study-confounding} = 0.5$			
	Mean	Var	Bias	EFF	Mean	Var	Bias	EFF
FE Two-stage SMA	0.4995	0.0062	-5e-04	101%	0.5692	0.0064	0.0692	63%
RE Two-stage SMA	0.4996	0.0063	-4e-04	100%	0.5445	0.007	0.0445	79%
FE Interaction MA	0.4996	0.0063	-4e-04	99%	0.4996	0.0063	-4e-04	112%
RE Interaction MA	0.4997	0.0064	-3e-04	98%	0.4997	0.0064	-3e-04	111%
Basic LMM model (1)	0.4996	0.0063	-4e-04	Ref.	0.5228	0.0066	0.0228	Ref.

Table 3: Simulating complete reported data with respectively six, ten and 20 studies. Results are with and without confounding. Efficiency (EFF) is calculated as the ratio between the mean squared error (MSE) of the basic LMM model compared to the specific models. The basic LMM model is used as the reference.

4.2 Incomplete reporting: Studies without all subgroups

We fit the basic LMM model with no bias adjustment

4.2.1 Four studies without all subgroups

$J = 6$	$\beta_{study-confounding} = 0$				$\beta_{study-confounding} = 0.5$			
	Mean	Var	Bias	EFF	Mean	Var	Bias	EFF
FE Two-stage SMA	0.5027	0.0353	0.0027	102%	0.7615	0.0425	0.2615	91%
RE Two-stage SMA	0.5027	0.0367	0.0027	98%	0.7515	0.0428	0.2515	95%
FE Interaction MA	0.5041	0.0714	0.0041	50%	0.5041	0.0714	0.0041	141%
RE Interaction MA	0.5038	0.0752	0.0038	48%	0.5038	0.0752	0.0038	134%
Basic LMM model (1)	0.5027	0.0361	0.0027	Ref.	0.7366	0.0444	0.2366	Ref.
$J = 10$	$\beta_{study-confounding} = 0$				$\beta_{study-confounding} = 0.5$			
	Mean	Var	Bias	EFF	Mean	Var	Bias	EFF
FE Two-stage SMA	0.498	0.0163	-0.002	101%	0.6654	0.0179	0.1654	71%
RE Two-stage SMA	0.4982	0.0167	-0.0018	99%	0.6486	0.0185	0.1486	79%
FE Interaction MA	0.4978	0.0215	-0.0022	77%	0.4978	0.0215	-0.0022	148%
RE Interaction MA	0.4981	0.0221	-0.0019	75%	0.4981	0.0221	-0.0019	145%
Basic LMM model (1)	0.4981	0.0165	-0.0019	Ref.	0.6107	0.0197	0.1107	Ref.
$J = 20$	$\beta_{study-confounding} = 0$				$\beta_{study-confounding} = 0.5$			
	Mean	Var	Bias	EFF	Mean	Var	Bias	EFF
FE Two-stage SMA	0.4999	0.007	-1e-04	101%	0.6158	0.0074	0.1158	52%
RE Two-stage SMA	0.4998	0.0071	-2e-04	100%	0.5943	0.0078	0.0943	65%
FE Interaction MA	0.5	0.008	0	89%	0.5	0.008	0	137%
RE Interaction MA	0.5	0.0081	0	87%	0.5	0.0081	0	135%
Basic LMM model (1)	0.4998	0.0071	-2e-04	Ref.	0.5544	0.008	0.0544	Ref.

Table 4: Simulating incomplete reported data with respectively six, ten and 20 studies. Four studies are missing a subgroup. Results are with and without confounding. Efficiency (EFF) is calculated as the ratio between the mean squared error (MSE) of the basic LMM model compared to the specific models. The basic LMM model is used as the reference.

4.2.2 50% of studies missing a subgroup

$J = 6$	$\beta_{study-confounding} = 0$				$\beta_{study-confounding} = 0.5$			
	Mean	Var	Bias	EFF	Mean	Var	Bias	EFF
FE Two-stage SMA	0.4997	0.0313	-3e-04	102%	0.6903	0.0359	0.1903	85%
RE Two-stage SMA	0.4996	0.0324	-4e-04	99%	0.6757	0.0369	0.1757	90%
FE Interaction MA	0.5009	0.0461	9e-04	69%	0.5009	0.0461	9e-04	132%
RE Interaction MA	0.5007	0.0481	7e-04	66%	0.5007	0.0481	7e-04	127%
Basic LMM model (1)	0.4998	0.032	-2e-04	Ref.	0.6492	0.0388	0.1492	Ref.
$J = 10$	$\beta_{study-confounding} = 0$				$\beta_{study-confounding} = 0.5$			
	Mean	Var	Bias	EFF	Mean	Var	Bias	EFF
FE Two-stage SMA	0.4996	0.0172	-4e-04	102%	0.7008	0.0192	0.2008	75%
RE Two-stage SMA	0.4996	0.0177	-4e-04	99%	0.6869	0.0198	0.1869	81%
FE Interaction MA	0.5	0.0265	0	66%	0.5	0.0265	0	168%
RE Interaction MA	0.5004	0.0271	4e-04	65%	0.5004	0.0271	4e-04	164%
Basic LMM model (1)	0.4996	0.0176	-4e-04	Ref.	0.6492	0.0221	0.1492	Ref.
$J = 20$	$\beta_{study-confounding} = 0$				$\beta_{study-confounding} = 0.5$			
	Mean	Var	Bias	EFF	Mean	Var	Bias	EFF
FE Two-stage SMA	0.4983	0.0088	-0.0017	101%	0.7077	0.0096	0.2077	61%
RE Two-stage SMA	0.4982	0.0089	-0.0018	100%	0.6925	0.0099	0.1925	68%
FE Interaction MA	0.4983	0.013	-0.0017	68%	0.4983	0.013	-0.0017	245%
RE Interaction MA	0.498	0.0132	-0.002	67%	0.498	0.0132	-0.002	241%
Basic LMM model (1)	0.4981	0.0089	-0.0019	Ref.	0.643	0.0115	0.143	Ref.

Table 5: Simulating incomplete reported data with respectively six, ten and 20 studies. 50% of studies are missing a subgroup. Results are with and without confounding. Efficiency (EFF) is calculated as the ratio between the mean squared error (MSE) of the basic LMM model compared to the specific models. The basic LMM model is used as the reference.

4.3 Studies not reporting subgroup specific treatment effects

We fit the basic LMM model with no bias adjustment

4.3.1 Four studies having pooled data

$J = 6$	$\beta_{study-confounding} = 0$				$\beta_{study-confounding} = 0.5$			
	Mean	Var	Bias	EFF	Mean	Var	Bias	EFF
FE Two-stage SMA	0.5021	0.0727	0.0021	70%	0.5361	0.0748	0.0361	137%
RE Two-stage SMA	0.5012	0.0775	0.0012	65%	0.5142	0.0806	0.0142	129%
FE Interaction MA	0.5019	0.0731	0.0019	69%	0.5019	0.0731	0.0019	143%
RE Interaction MA	0.5009	0.0782	9e-04	65%	0.5009	0.0782	9e-04	133%
Basic LMM model (1)	0.467	0.0495	-0.033	Ref.	0.6674	0.0763	0.1674	Ref.
$J = 10$	$\beta_{study-confounding} = 0$				$\beta_{study-confounding} = 0.5$			
	Mean	Var	Bias	EFF	Mean	Var	Bias	EFF
FE Two-stage SMA	0.4992	0.0222	-8e-04	86%	0.5605	0.023	0.0605	104%
RE Two-stage SMA	0.4991	0.0228	-9e-04	83%	0.5366	0.0247	0.0366	107%
FE Interaction MA	0.4991	0.0225	-9e-04	84%	0.4991	0.0225	-9e-04	123%
RE Interaction MA	0.499	0.0232	-0.001	82%	0.499	0.0232	-0.001	120%
Basic LMM model (1)	0.4839	0.0187	-0.0161	Ref.	0.5632	0.0238	0.0632	Ref.
$J = 20$	$\beta_{study-confounding} = 0$				$\beta_{study-confounding} = 0.5$			
	Mean	Var	Bias	EFF	Mean	Var	Bias	EFF
FE Two-stage SMA	0.5005	0.0078	5e-04	93%	0.5693	0.0081	0.0693	73%
RE Two-stage SMA	0.5005	0.0079	5e-04	92%	0.5447	0.0087	0.0447	89%
FE Interaction MA	0.5006	0.008	6e-04	91%	0.5006	0.008	6e-04	119%
RE Interaction MA	0.5007	0.0081	7e-04	90%	0.5007	0.0081	7e-04	118%
Basic LMM model (1)	0.4931	0.0072	-0.0069	Ref.	0.5367	0.0081	0.0367	Ref.

Table 6: Simulating incomplete reported data with respectively six, ten and 20 studies. Four studies are not reporting subgroup specific treatment effects. Results are with and without confounding. Efficiency (EFF) is calculated as the ratio between the mean squared error (MSE) of the basic LMM model compared to the specific models. The basic LMM model is used as the reference.

4.3.2 50% of studies missing a subgroup

$J = 6$	$\beta_{study-confounding} = 0$				$\beta_{study-confounding} = 0.5$			
	Mean	Var	Bias	EFF	Mean	Var	Bias	EFF
FE Two-stage SMA	0.4994	0.0459	-6e-04	81%	0.5465	0.0478	0.0465	116%
RE Two-stage SMA	0.4984	0.0481	-0.0016	77%	0.5227	0.051	0.0227	112%
FE Interaction MA	0.4992	0.0464	-8e-04	80%	0.4992	0.0464	-8e-04	124%
RE Interaction MA	0.4984	0.0486	-0.0016	76%	0.4984	0.0486	-0.0016	119%
Basic LMM model (1)	0.4793	0.0367	-0.0207	Ref.	0.5878	0.05	0.0878	Ref.
$J = 10$	$\beta_{study-confounding} = 0$				$\beta_{study-confounding} = 0.5$			
	Mean	Var	Bias	EFF	Mean	Var	Bias	EFF
FE Two-stage SMA	0.499	0.0269	-0.001	81%	0.5576	0.028	0.0576	116%
RE Two-stage SMA	0.4989	0.0278	-0.0011	78%	0.5342	0.0299	0.0342	117%
FE Interaction MA	0.499	0.0272	-0.001	80%	0.499	0.0272	-0.001	133%
RE Interaction MA	0.499	0.0281	-0.001	77%	0.499	0.0281	-0.001	129%
Basic LMM model (1)	0.4789	0.0213	-0.0211	Ref.	0.5866	0.0288	0.0866	Ref.
$J = 20$	$\beta_{study-confounding} = 0$				$\beta_{study-confounding} = 0.5$			
	Mean	Var	Bias	EFF	Mean	Var	Bias	EFF
FE Two-stage SMA	0.4992	0.0127	-8e-04	84%	0.5649	0.0132	0.0649	110%
RE Two-stage SMA	0.4991	0.0129	-9e-04	83%	0.5409	0.014	0.0409	123%
FE Interaction MA	0.4995	0.0129	-5e-04	83%	0.4995	0.0129	-5e-04	148%
RE Interaction MA	0.4995	0.0131	-5e-04	81%	0.4995	0.0131	-5e-04	146%
Basic LMM model (1)	0.4787	0.0102	-0.0213	Ref.	0.5743	0.0137	0.0743	Ref.

Table 7: Simulating incomplete reported data with respectively six, ten and 20 studies. 50% of studies do not report subgroup specific treatment effects. Results are with and without confounding. Efficiency (EFF) is calculated as the ratio between the mean squared error (MSE) of the basic LMM model compared to the specific models. The basic LMM model is used as the reference.

8.2 Manuscript II

RTSA: An R package for the updated version of Trial Sequential Analysis

Anne Lyngholm Soerensen, Markus Harboe Olsen, Theis Lange & Christian Gluud

Details: Submitted to *Journal of Statistical Software* in 2023.




Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II.

doi: 10.18637/jss.v000.i00

RTSA: An R Package for the Updated Version of Trial Sequential Analysis

Anne Lyngholm Sørensen 
University of Copenhagen
Copenhagen Trial Unit, Rigshospitalet

Markus Harboe Olsen
Copenhagen Trial Unit, Rigshospitalet

Theis Lange
University of Copenhagen

Christian Gluud
Copenhagen Trial Unit, Rigshospitalet
University of Southern Denmark

Abstract

The R Trial Sequential Analysis (**RTSA**) package provides a set-up for trial sequential analysis in R. Trial sequential analysis is an adaptation of group sequential methods for randomised clinical trials to meta-analyses. The functions included in the package are meant to aid the user in creating either prospective or retrospective end-to-end sequential meta-analyses. Starting from the meta-analysis itself till adopting or investigating the effect of a sequential testing regime, the user will be provided with guidance about the validity of the evidence, interpretation of the inference, and more. As many considerations are required to do a proper analysis of a sequentially updated meta-analysis, the package includes several vignettes to support the users of the package. The sequential methods for meta-analyses in the **RTSA** package stems from Trial Sequential Analysis, software implemented in Java, and can be considered an extended version of the software originally implemented as stand-alone. The R package **RTSA** offers a larger library of methods compared to Trial Sequential Analysis in Java, including new gold standards within sequentially updated meta-analysis.

Keywords: Trial Sequential Analysis, Sequential meta-analysis, Retrospective meta-analysis, Prospective meta-analysis, Trial Sequential Analysis in R.

1. Introduction

Randomized trials are often repeated in different settings and with only minor differences. A

meta-analysis is the statistical synthesis of these trials' results, boiling down the different effect estimates to one intervention effect estimate. This pooled estimate aims to encapsulate the general direction, size, and variability of the individual trials' effect estimates. Besides getting one overall estimate of the effect, a potential benefit of the meta-analysis is the increase in power [Deeks *et al.* \(2019\)](#). When new trials emerge, and more evidence is created, the meta-analysis can be updated to potentially add more power and also to update the intervention effect estimate. However, many meta-analyses are updated without adjusting for multiplicity.

R has an extensive library of meta-analysis packages, some of the most popular being: **metafor** [Viechtbauer \(2010\)](#), **meta** [Schwarzer *et al.* \(2015\)](#), and **dmetar** [Harrer *et al.* \(2019\)](#). Evidence from a well-conducted meta-analysis is considered to be one of the best sources for assessment of intervention effects [Ioannidis \(2022\)](#). It also captures trends by synthesizing results as evidence accumulates over time as described in a Living Systematic Review [Simmonds *et al.* \(2017\)](#). In the effort to reduce research waste, some journals require at least one meta-analysis, if not a systematic review with a meta-analysis, before submitting a randomised clinical trial [Chalmers *et al.* \(2014\)](#). However, a meta-analysis can have problems with the validity of its statistical analyses. By being updated over time, as it will in many applications, it is difficult to control type-I- and type-II-errors. **RTSA** aims to provide statistical methods and guidance to the statistical analysis when a meta-analysis becomes sequential.

Updating a meta-analysis and thereby repeating the same hypothesis test will inflate the type-I-error, hence the risk of false rejection of the null hypothesis is increasing above the specified level often set to 2.5% or 5% [Wetterslev *et al.* \(2017\)](#) [Pogue and Yusuf \(1997\)](#) [Wetterslev *et al.* \(2008\)](#) [Imberger *et al.* \(2015\)](#) [Imberger *et al.* \(2016\)](#). Several methods, most of them inspired by group sequential methods in clinical trials, have emerged to minimize the inflation of the type-I-error in cumulative meta-analysis. Group sequential methods allow single trials to analyse the null hypothesis multiple times as the trial data accumulates during the trial while still preserving the type-I-error. The trial can stop early if the hypothesis can be rejected. Each analysis during a single trial is called an interim analysis, while the analysis carried out when all data has been accumulated (if the trial has not stopped early) is called the final analysis. Meta-analysis can be adapted to fit into a group sequential testing regime by having interim analyses, e.g. per newly added trial. Besides allowing for more control of the type-I-error, another reason to impose a sequential design to cumulative meta-analysis is to be able to stop recruitment of new trials early. Just as in sequential methods for a single trial, stopping early for findings of superiority, inferiority, or futility can be achieved. Another key point of using sequential meta-analysis is the ability to derive the power of the analysis together with the required meta-analytic sample size calculation. Many meta-analyses are underpowered even though the authors are not made aware [Wetterslev *et al.* \(2008\)](#).

There are multiple methods for sequential meta-analysis. Among these are Bayesian sequential meta-analysis including semi-Bayesian [Turner *et al.* \(2014\)](#), fully-Bayesian [Spence *et al.* \(2016\)](#) methods, and frequentist sequential meta-analysis such as sequential meta-analysis by Whitehead [Whitehead \(2002\)](#), and Trial Sequential Analysis (TSA) [Wetterslev *et al.* \(2008\)](#). None of these methods are currently implemented in R. The user must instead first use a package to conduct a meta-analysis and then try to fit that meta-analysis into one of the group sequential packages developed for clinical trials. These include **gsDesign** [Anderson \(2022\)](#) and **rpact** [Wassmer and Pahlke \(2022\)](#). **RTSA** intends to fill this exact gap with an updated implementation of TSA.

TSA is a software created for sequential meta-analysis which, until now, only has been imple-

mented as stand-alone and written in Java [Wetterslev et al. \(2008\)](#) [Thorlund et al. \(2011\)](#). As gold standards of doing both meta-analysis and sequential meta-analysis develop over time, the Java implementation does not include all current best-practices. Therefore, the **RTSA** implementation includes several new features compared to the original Java implementation. **RTSA** is easier to update with new features or best-practices and adds transparency for a larger population of users of the code, since the code is publicly available. The assumptions of the methods implemented can now be more easily investigated and tested by, e.g. simulation studies which has not been possible to do without hard coding until now. Another great advantage of the **RTSA** package is the vignettes which can aid practitioners in using the software in the best possible way.

This paper serves as an introduction to the main functionalities of **RTSA** and is structured as follows. A condensed version of the background theory surrounding meta-analysis and group sequential methods is given in Section 2. Section 3 describes some of the most important considerations to be made before using sequential meta-analysis. Section 4 describes the package structure and how to use the main function `RTSA()`. `RTSA()` consists of four subfunctions: `metaanalysis()`, `ris()`, `boundaries()` and `inference()`. The most important of the subfunctions will be described in Section 5. Finally, in Section 6, we discuss the main contributions of the package including the future plans for **RTSA**.

2. Methods

We start by describing some background theory on meta-analysis and sequential meta-analysis.

2.1. Meta-analysis

Consider K trials all investigating an identical or comparable hypothesis. Each study k , where $k = 1, \dots, K$, provides an estimated effect size $\hat{\theta}_k$ and standard error s_k . A meta-analysis synthesizes the effect sizes into one pooled effect size by taking a weighted average of the observed effect sizes. A common choice of measure for the weights is the inverse of the variance of the trial estimate. The weights which we denote w_k , the pooled effect size $\hat{\theta}$ and the variance of $\hat{\theta}$ can then be calculated by:

$$w_k = \frac{1}{s_k^2}, \quad \hat{\theta} = \frac{\sum_k w_k \hat{\theta}_k}{\sum_k w_k}, \quad \text{and} \quad \text{var}(\hat{\theta}) = \frac{1}{\sum_k w_k}. \quad (1)$$

Other methods for weighting are Mantel-Haenszel (MH) methods which are appropriate to use for binary outcome data when the number of events and participants are small and there is a large number of trials [Deeks et al. \(2019\)](#). The MH methods can be used for odds ratios (OR), relative risks (RR), or risk differences (RD).

The model in (1) assumes that there is one true effect θ , which we are trying to estimate and that $\hat{\theta}_k \sim \mathcal{N}(\theta, \sigma_k^2)$. Such a meta-analysis is called a fixed-effect or common-effect meta-analysis. If the trials included does not apply almost identical trial designs or it is believed that the trials for other means are different from each other, but are still comparable, it might be more appropriate to use a random-effects meta-analysis. Here we assume instead that $\hat{\theta}_k \sim \mathcal{N}(\theta_k, \sigma_k^2)$ with $\theta_k \sim \mathcal{N}(\theta, \tau^2)$. Hence the assumption of one true effect is no longer true. Instead each trial is expected to have its own true effect and the meta-analysed estimate is now the mean of a distribution of trial effects. The random-effects meta-analysis is

conducted by including a term for heterogeneity, τ^2 , to the weights. τ^2 describes the between-trial variation. The assumption of a random-effects meta-analysis changes the pooled effect estimate and its estimated variance from (1) into:

$$w_k^R = \frac{1}{s_k^2 + \hat{\tau}^2}, \quad \hat{\theta}^R = \frac{\sum_k w_k^R \hat{\theta}_k}{\sum_k w_k^R}, \quad \text{and} \quad \text{var}(\hat{\theta}^R) = \frac{1}{\sum_k w_k^R}. \quad (2)$$

Here the common method for estimating the variance in the random-effects meta-analysis is done using an estimator by DerSimonian-Laird [DerSimonian and Laird \(1986\)](#). An adjustment of the variance estimate which assumes a t distribution instead of a normal distribution is the Hartung-Knapp-Sidik-Jonkman adjustment, which has been shown to be more stable when having a smaller number of trials [IntHout et al. \(2014\)](#).

When the heterogeneity, τ^2 , has been estimated there exists several methods to quantify the size of the heterogeneity relative to the noise. Two of them are inconsistency (I^2) introduced by [Higgins and Thompson \(2002\)](#) [Higgins \(2003\)](#) and diversity (D^2) introduced by [Wetterslev et al. \(2009\)](#):

$$I^2 = \frac{\tau^2}{\tau^2 + \sigma_M^2} \quad \text{and} \quad D^2 = \frac{\tau^2}{\tau^2 + \sigma_D^2}. \quad (3)$$

Both aim at estimating the proportion of between-trial variation τ^2 of the total variation with differing assumptions about the within-trial error σ^2 , calculated by σ_M^2 [Higgins and Thompson \(2002\)](#) or σ_D^2 [Wetterslev et al. \(2009\)](#).

Sample and trial size calculation

An estimate of the sample size of the meta-analysis is required to perform a sequential meta-analysis using TSA. For fixed-effect meta-analysis, only a minimum sample size is required to achieve a specific power. Sample size is labelled *required information size* (RIS) in TSA [Wetterslev et al. \(2009\)](#). By the assumption of normality of the pooled effect estimate, the RIS of a meta-analysis under the fixed-effect model can be calculated by:

$$RIS = 4 \cdot (z_{1-\alpha/side} + z_\beta)^2 \cdot \frac{\nu}{\theta^2}. \quad (4)$$

Where z_x is the x 'th quantile of the standard normal distribution, α is the type-I-error, $side$ is the side of the test (one or two-sided), β is the type-II-error ($1 - \beta$ is the power), ν is the expected variance and θ is the minimal clinical value of interest [Pogue and Yusuf \(1997\)](#). For binary data are $\nu = (1 - p_0)p_0$, $p_0 = (p_I + p_C)/2$, and $\theta = p_C - p_I$, with p_I being the probability for event in the intervention group and p_C being the probability for event in the control group. θ is either the minimum clinical relevant RD or the log of the minimum clinical relevant OR or RR.

Under the assumption of a random-effects meta-analysis, it is necessary to have both a minimum number of trials as well as participants to achieve a specific power [Kulinskaya and Wood \(2013\)](#). Given a value of the heterogeneity τ it is possible to calculate the minimum required number of trials needed using:

$$\frac{\tilde{\theta}}{\sqrt{\text{Var}(\tilde{\theta})}} = z_{1-\alpha/side} + z_{1-\beta}, \quad \text{where} \quad \text{Var}(\tilde{\theta}) = \left(\sum_k \frac{1}{2 \cdot s_k^2/n_k + \tau^2} \right)^{-1}, \quad (5)$$

where n_k is the number of participants in trial k . Then, we will have the defined power, $1 - \beta$ when:

$$\tau^2 < \frac{\theta \cdot K}{\left(z_{1-\alpha/side} + z_{1-\beta}\right)^2}.$$

here K is the number of trials. With an estimate of K , we can calculate the number of participants per trial:

$$n_k = \frac{2 \cdot \sigma^2}{\frac{\theta \cdot K}{\left(z_{1-\alpha/side} + z_{1-\beta}\right)^2} - \tau^2} \quad (6)$$

The formulas are all derived from [Kulinskaya and Wood \(2013\)](#). Note that K is a minimum number of trials required. Any choice of $K' > K$ will also achieve the specified level of power while affecting the number of required participants per trial. Often the total participants requirement will be less for larger K , however, too large a K will result in very small trials removing the value of the single trials.

Under the assumption of a random-effects model, other suggestions have been made to calculate the sample size of a meta-analysis using inconsistency I^2 or diversity D^2 :

$$RIS_{D^2} = \frac{1}{1 - D^2} \cdot 4 \cdot (z_{1-\alpha/2} + z_{1-\beta})^2 \cdot \frac{\nu}{\theta^2}.$$

$$RIS_{I^2} = \frac{1}{1 - D^2} \cdot 4 \cdot (z_{1-\alpha/2} + z_{\beta})^2 \cdot \frac{\nu}{\theta^2}.$$

It has been shown that the power requirements are not guaranteed to be met using the diversity adjustment [Kulinskaya and Wood \(2013\)](#). As the inconsistency adjustment leads to smaller or equal sample sizes to the diversity adjusted sample size [Wetterslev *et al.* \(2009\)](#), we expect that the power requirements are also not guaranteed when adjusting by inconsistency.

Note that the meta-analysis sample size requirement will change when the meta-analysis is using a sequential design. A sequential meta-analysis using a group sequential design will have a larger required information size than those calculated with formulas (4) for fixed-effect meta-analysis or (6) for random-effects meta-analysis. To achieve the wanted level of power when using a sequential design, we need to scale the RIS calculated from (4) or (6). We will see examples of this in the next subsection 2.2, and in subsection 5.3 we will compare the sample size of a sequential meta-analysis with and without a sequential design.

2.2. Sequential meta-analysis

Sequential meta-analysis, as in TSA, is based almost entirely on sequential methods for clinical trials [Wetterslev *et al.* \(2017\)](#). If familiar with group sequential designs, it will be clear that sequential meta-analysis is an adaptation of the key elements of group sequential designs. The following subsections will describe some of the key elements of group sequential methods but will not go into depth about the technical details. [Jennison and Turnbull \(1999\)](#) and [Wassmer and Brannath \(2016\)](#) are great resources for the theory behind group sequential methods.

Sequential testing

Consider a group of research centers deciding to conduct K trials with the purpose of sequentially combining them in meta-analyses – i.e. planning a prospective sequential meta-analysis.

After each trial finishes the results of the finished trials are combined in a meta-analysis resulting in $K - 1 = M$ potential meta-analyses. At each interim meta-analysis m , $m = 1, \dots, M$, the null hypothesis of interest is tested. Here $m = 1$ corresponds to the first two trials being meta-analysed, $m = 2$ corresponds to the first three trials being meta-analysed, etc. At each interim, the test statistic is calculated by:

$$Z_m = \frac{\hat{\theta}^{(m)}}{\sqrt{\text{var}(\hat{\theta}^{(m)})}}, \quad (7)$$

where $\hat{\theta}^{(m)}$ is the pooled effect estimated from meta-analysis m . At each interim m , the test statistic is evaluated for whether the sequential meta-analysis can stop. The testing thresholds for stopping are dependent on the time of testing. Let b_m for $m = 1, \dots, M$ be a testing threshold/boundary for the sequential meta-analysis m . Consider that we are interested in stopping for efficacy or harm. Let b_m be the stopping boundaries for efficacy and $-b_m$ for harm. The stopping boundaries are designed such that the type-I-error is protected. As the design is two-sided, we have symmetric stopping boundaries b_m and $-b_m$ in this scenario. A sequential testing scheme is then:

For $m = 1, \dots, M - 1$:	
if $Z_m < -b_m$ or $Z_m > b_m$	stop for efficacy or harm
else	continue sequential meta-analysis
For $m = M$:	
if $Z_M < -b_M$ or $Z_M > b_M$	stop for efficacy or harm
else	stop, the null hypothesis could not be rejected.

How the boundaries should be calculated depends on the distribution of $\mathbf{Z} = (Z_1, Z_2, \dots, Z_M)$, the timing of the trials in the meta-analysis relative to the required information size, and how we split the type-I-error and potentially type-II-error during the sequential meta-analysis. Error-spending functions can be used for splitting the errors across the sequential meta-analysis. If the function splits the type-I-error, we can call them α -spending functions.

Error spending functions

The control of type-I-error is achieved by splitting the risk of a type-I-error across the analyses (interims and final). A simple example is to split a type-I-error of 0.05 into two, so 0.025 is spent at the interim analysis and 0.025 is spent at the final analysis. There are a variety of established so-called α -spending functions. These functions determine the amount of α spent across the sequential meta-analysis. α -spending functions allow for flexibility in the design by not requiring to know the exact timing of future interim analyses at the time of the current interim. Some examples of error-spending functions are Lan and DeMets' versions of O'Brien-Fleming's and Pocock's boundaries [DeMets and Lan \(1994\)](#), the Hwang-Shih-DeCani's error spending function [Hwang *et al.* \(1990\)](#), and the power family [Wang and Tsiatis \(1987\)](#), [Emerson and Fleming \(1989\)](#), and [Pampallona and Tsiatis \(1994\)](#).

The functions used for α -spending can also be used for β -spending, where the type-II-error is split across the interim and final analyses making it possible to stop the meta-analysis for futility. Futility stopping boundaries describe when it is unlikely that the null hypothesis

will be rejected. With the use of futility boundaries it is possible to stop the sequential meta-analysis early due to a slim chance of rejecting the null hypothesis.

There are two options when using β -spending as futility boundaries. They are defined as either binding or non-binding. When using binding futility boundaries, it is strictly assumed that one will stop the meta-analysis when entering the futility area (crossing a β -spending boundary). For non-binding futility, it is assumed that one does not stop the meta-analysis when entering the futility area. More information about binding compared to non-binding futility can be found in the Futility boundaries vignette.

For examples of α -spending and β -spending boundaries see Figure 1. Here the red lines are the α -spending boundaries, the blue lines are the β -spending boundaries, and the green lines are the naive testing boundaries. Crossing a red line means that the null-hypothesis can be safely rejected without risking inflation of the type-I-error when conducting multiple hypothesis tests if the meta-analysis is prospective. Figure 1 also shows the increase in required information size at the top right corner, where we find that the first design (left plot) increases RIS with 2%. The second design (right plot) increases RIS with 26%. The increase in RIS is due to the sequential design. To reach the wanted level of power under multiple testing in a group sequential design, the sample size increases compared to a design where we only test the hypothesis once.

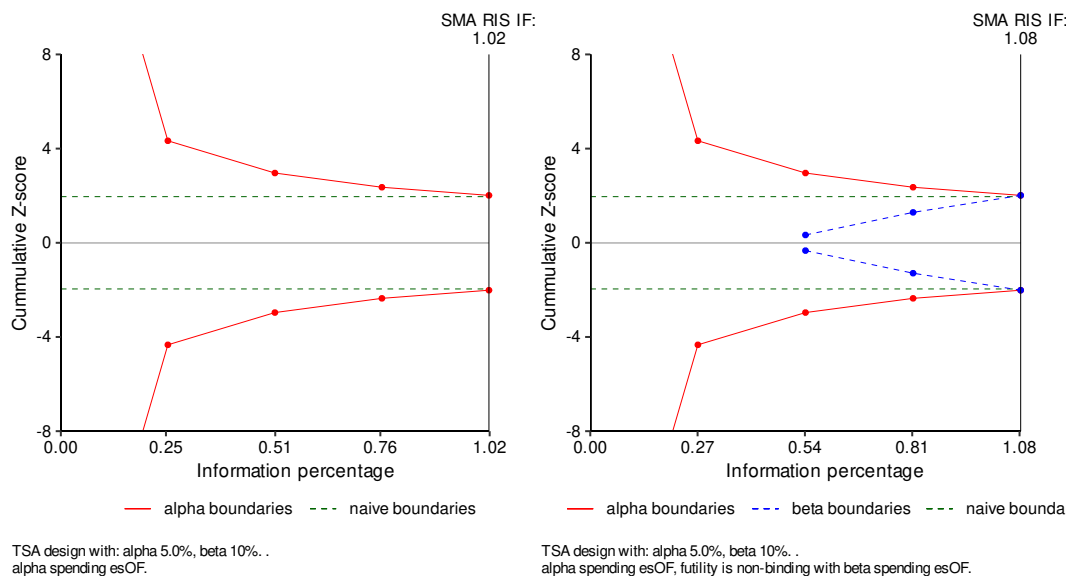


Figure 1: Examples of α - and β -spending boundaries. Left plot: Two-sided design with α -spending boundaries (red). Right plot: Two-sided design with α -spending boundaries (red) and β -spending boundaries set to non-binding futility (blue). All α - and β -boundaries are computed using error spending function Lan and DeMets' version of O'Brien-Fleming stopping boundaries. The sequential designs are created using the `plot()` function on a `boundaries` object from the **RTSA** package. SMA: sequential meta-analysis. RIS: required information size.

The testing boundaries are updated as the trials are finished and incorporated in the sequential

meta-analysis, as it is often not possible to recruit exactly the required number of participants. Depending on how large the difference is between the assumptions of the design and observed data, the sequential meta-analysis might be over- or underpowered. Furthermore, inference such as confidence intervals can be calculated at each interim as the sequential meta-analysis is updated.

Inference at interim analyses and final analysis

The validity of the naive point estimates, confidence intervals, and p-values are affected by the sequential design. Computing the metrics naively as done in non-sequential meta-analyses results in biased estimates if a sequential design is used. There is a selection of methods to use for adjusting the estimates when having interim analyses in the design. In the scenario of prospective meta-analyses (see Subsection 2.3), we should use one method while the meta-analysis is still running and another for when it has stopped. The following two methods can be used for each scenario, respectively:

- *TSA-adjusted method*: Computes confidence intervals for sequential meta-analysis with correct sequence coverage when the meta-analysis has not stopped. This inference adjustment method is used for continued sequential meta-analyses. These confidence intervals will, however, be too conservative if the sequential meta-analysis has stopped, either by crossing a stopping boundary or reaching the required information size, and should not be used in this scenario.
- *Sample space ordered methods*: Computes an unbiased median point estimate, confidence interval, and p-value for a stopped sequential meta-analyses. Can only be used for meta-analyses that either stopped early or continued to the final analysis. There are several ways to order the sample space. One of the most popular being the stage-wise ordering proposed by [Armitage \(1957\)](#) which is implemented in **RTSA**.

TSA-adjusted confidence intervals are identical to what is known as repeated confidence intervals by [Jennison and Turnbull \(1984\)](#). More information about repeated inference and the stage-wise ordering can be found in the book by [Jennison and Turnbull \(1999\)](#) and the book by [Wassmer and Brannath \(2016\)](#).

In the case of retrospective meta-analyses and living systematic reviews (presented below), there is no guarantee for the validity of the inference. A similar practice of using the aforementioned inference methods can be applied, but one must be conservative when interpreting the results.

2.3. Sequential meta-analysis by design (prospective) or by circumstance (retrospective)

The functions in **RTSA** depend on whether the meta-analysis is prospective or retrospective. The type is defined by the timing of the trials relative to the planning of the meta-analysis. A prospective meta-analysis requires that all trials to be included in the meta-analysis are yet to be performed or at least that the trials' results are completely unknown when deciding to conduct the meta-analysis (see subsection 2.2.1). The meta-analysis is retrospective if all or some trials' results are published or known. This is the most common situation. Regardless of

which type of meta-analysis is considered, the decision of conducting a traditional or sequential meta-analysis should be accompanied by a predefined protocol and statistical analysis plan (SAP). The sequential design can be chosen for the prospective meta-analysis which will ensure correct control of type-I-error even if there are multiple interim analyses of the meta-analysis. **RTSA** can provide designs for prospective sequential meta-analyses including meta-analysis sample size calculations. For more information about prospective meta-analysis see [Seidler *et al.* \(2019\)](#).

For most meta-analyses some or all of the results of the trials will be known, and there might even be a prior meta-analysis which one now wishes to update. This means that the meta-analysis is retrospective and caution should be exercised during the interpretation of the results of a sequential meta-analysis as they should with a naive meta-analysis on the same data. Knowledge of trial results or an existing meta-analysis could introduce sequential decision bias (the decision to make a new trial is conditional on the result of the known trials or meta-analysis) or sequential design bias (the design of new trials are based on results from earlier trials) in the meta-analysis [Kulinskaya *et al.* \(2015\)](#). Sequential decision bias will cause promising meta-analyses to be more likely to be continued while less favorable meta-analyses will more likely be discontinued. This creates an upwards bias, an overestimation of the point estimate, on average. Regardless of the potential bias of retrospective sequential meta-analysis, a sequential meta-analysis will exert more control over type-I-errors than the naive meta-analysis, as well as offering an area of futility.

3. Using RTSA

The main function of **RTSA** is the `RTSA()` function which can compute a complete sequential meta-analysis end-to-end. This includes several elements such as a naive meta-analysis, a Trial Sequential Analysis (sequential boundaries and interim inference), information about the sample size, and inference of the cumulative Trial Sequential Analysis at the final analysis. The elements are divided into four sub-functions which all are used by `RTSA()`. Each of the sub-elements of `RTSA()` can be run individually:

- `metaanalysis()` computing a meta-analysis with naive tests,
- `ris()` conducts a sample and trial size calculation for retrospective or prospective meta-analysis used at the planning stage,
- `boundaries()` calculating testing boundaries and a measure for scaling the sample size for correct power for the sequential meta-analysis, and,
- `inference()` computing naive and conditional inference based on the stage of the sequential meta-analysis.

Each of these can be called individually, if only specific parts of the sequential meta-analysis are of interest.

The use of `RTSA()` depends on the aim and whether the sequential meta-analysis is prospective or retrospective. If the sequential meta-analysis is prospective, two scenarios are possible:

1. *Planning*: Plan a prospective sequential meta-analysis. This must be done prior the conduct of all the trials which are to be included or at least prior to analysis and publication of results of one or more of the trials.
2. *Analyse prospective meta-analysis*: Compute a naive meta-analysis and TSA with conditional inference using the planned design as the data on finished trials accumulate.

If the sequential meta-analysis is retrospective only one option is possible:

3. *Analyse retrospective meta-analysis*: Compute a naive meta-analysis, TSA and conditional inference with a retrospectively created design as the trials used in the meta-analysis accumulate.

Table 1 and 2 describes the arguments used in the `RTSA()` function depending on the whether the function is used for design or analysis as described above. We will now describe the `RTSA()` function in greater detail using toy-data for prospective sequential meta-analysis and a real data example for retrospective sequential meta-analysis.

3.1. Planning a prospective sequential meta-analysis

In planning a sequential meta-analysis, `RTSA()` uses the sub-functions `ris()` and `boundaries()`. The subfunctions are in this scenario not dependent on each other as shown on Figure 2.

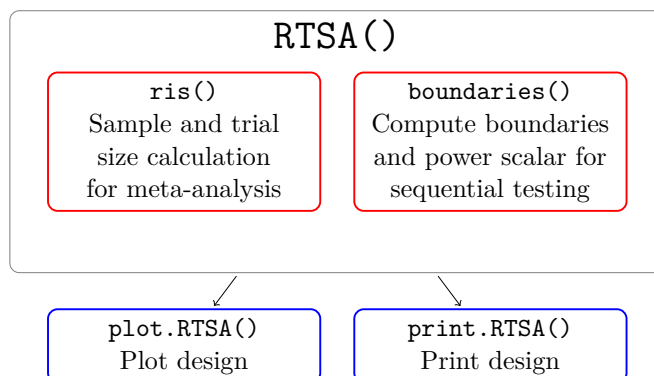


Figure 2: Subfunctions used in `RTSA()` and types of output when `RTSA()` is used for planning a prospective sequential meta-analysis.

To use the function for planning, parameters for a sample and trial size calculation are required as well as an estimate of the cumulative timing of the trials. The main arguments for a design in the `RTSA()` are:

```
RTSA(type = "design", outcome, timing, mc, ...)
```

The argument `type = "design"` is used for specifying that the `RTSA()` function is used for planning a prospective sequential meta-analysis. Furthermore, the design needs to be either fixed-effect or random-effects meta-analysis. If a random-effects meta-analysis is anticipated

Argument	Description	Prospective SMA		Retrospective SMA
		Design	Analysis	Analysis
<i>Mandatory arguments for all purposes</i>				
type	Specify whether the <code>RTSA()</code> function is used for design or analysis	Required (design)	Required (analysis)	Required (analysis)
<i>Design arguments</i>				
outcome	Outcome metric. Options are "MD" (mean difference), "RR" (risk ratio), "OR" (odds ratio) and "RD" (risk difference).	Required	Taken from design	Required
side	Whether the test is one- or two-sided	Required	Taken from design	Required
alpha	Level of type-I-error	Required	Taken from design	Required
beta	Level of type-II-error	Required	Taken from design	Required
futility	Choice of futility boundaires	Optional	Taken from design	Optional
es_alpha	Error spending function used for α -spending	Required	Taken from design	Required
es_beta	Error spending function used for β -spending	Optional	Taken from design	Optional
timing	Timing of trials	Required	Taken from design	Inapplicable
<i>Analysis arguments</i>				
data	Data for meta-analysis	Inapplicable	Required	Required
design	Design for prospective meta-analysis	Inapplicable	Required	Inapplicable
<i>Meta-analysis arguments</i>				
weights	Weight method for meta-analysis	Inapplicable	Optional	Optional
re_method	Method for estimating the variance in a random-effects model	Inapplicable	Optional	Optional
tau_ci_method	Method for calculating the confidence interval for the estimate of heterogeneity	Inapplicable	Optional	Optional

Table 1: Arguments used in the `RTSA()` function

Argument	Description	Prospective SMA		Retrospective SMA
		Design	Analysis	Analysis
<i>Specific for sample size calculation</i>				
<code>fixed</code>	Whether the meta-analysis is fixed-effect or random-effects	Required	Taken from design	Required
<code>mc</code>	Minimum clinical relevant value	Required	Taken from design	Required
<code>sd_mc</code>	Expected standard deviation of outcome	Depends on outcome	Taken from design	Depends on outcome
<code>pc</code>	Expected probability of event in control group	Depends on outcome	Taken from design	Depends on outcome
<code>random_adj</code>	The sample size adjustment based on presence of heterogeneity. Options are "D2" (Diversity), "I2" (Inconsistency) and "tau2" (the heterogeneity estimate). Default is "tau2".	Optional	Optional	Optional
<i>Miscellaneous or less used arguments</i>				
...				
See ?RTSA for more arguments				

Table 2: Additional arguments used in the RTSA() function

one must provide a best guess of the heterogeneity expressed by τ^2 , but options to use inconsistency I^2 or diversity D^2 are possible. For the required arguments see Table 1 and 2.

Suppose we are interested in a two-sided test design with binary outcome data using RR as the outcome metric. The event rate in the control group is expected to be 10%. A relative risk reduction (RRR) of 20% (RR of 0.8) is the minimum relevant reduction in risk. We want to have 90% power and a type-I-error of 5%. Four trial centers have agreed on a sequential meta-analysis, where each center will contribute with a quarter of the participants, a plausible timing is then $c(0.25, 0.5, 0.75, 1)$. Thus, interim analyses are planned at 25%, 50%, and 75% of participants accumulated. They plan to fit a fixed-effect meta-analysis, want binding futility and both the α - and β -spending functions are of type Lan and DeMets' version of O'Brien-Fleming stopping boundaries. A design can then be made as:

```
R> design_RTSA <- RTSA(type = "design", outcome = "RR", pC = 0.1, mc = 0.8,
+                       side = 2, fixed = TRUE, alpha = 0.05, beta = 0.1,
+                       timing = c(0.25, 0.5, 0.75, 1.0), es_alpha = "esOF",
+                       futility = "binding",
+                       es_beta = "esOF")
```

Here the arguments `p0`, `mc`, `side`, `alpha`, `beta`, and `side` are used for the sample size calculation by the `ris()` function and `alpha`, `beta`, `timing`, `es_alpha`, `futility`, and `es_beta` are used for calculating the testing boundaries by the `boundaries()`. A thorough introduction to the subfunctions is provided in Section 4.

The output of `RTSA()` describes that it was used for design and includes the boundaries for the sequential design as well as a sample size calculation with and without an adjustment for the sequential design:

```
R> design_RTSA
```

Design with Trial Sequential Analysis was computed with the following settings:

Boundaries for a 2-sided design with a type-I-error of 0.05, and type-II-error of 0.1.

Futility is set to: binding. Alpha-spending function: esOF.

Beta-spending function: esOF.

The required information size is not adjusted by heterogeneity. The required information size is further increased with 5 percent due to the sequential design. The total required information size is 9064.

Timing, and boundaries:

sma_timing	upper	lower	fut_upper	fut_lower
0.263	4.333	-4.333	NA	NA
0.527	2.963	-2.963	0.299	-0.299
0.790	2.359	-2.359	1.251	-1.251
1.053	1.963	-1.963	1.963	-1.963

`sma_timing` is the ratio of the required sample for a sequential meta-analysis

to a non-sequential meta-analysis sample size.

Sample size calculation for standard meta-analysis:

This is a prospective meta-analysis sample size calculation.

The sample size calculation assumes a 2-sided test, equal group sizes, a type-I-error of 0.05 and a type-II-error of 0.1.

The minimum clinical relevant value is set to: 0.8 for outcome metric RR.

Additional parameters for sample size are:

Probability of event in the control group: 0.1.

Fixed-effect required information size:

8606 participants in total.

For more information about the sample size calculation see vignette:

'Calculating required sample size and required number of trials'.

Sample size calculation for sequential meta-analysis:

Fixed-effect: 9064 participants.

Please note the following warnings:

- The RTSA function is used for design. Boundaries are computed but sequential inference will not be calculated. Use the `metaanalysis()` function if interested in meta-analysis results.

It is seen from the output that the sequential meta-analysis timing requires 1.053 RIS to reach the wanted level of power. The RIS calculation gives 8606 participants but the required sample size the sequential meta-analysis is $8606 \cdot 1.053 \approx 9064$. The design can be presented via `print()` as just shown or as a plot using `plot()`. The boundaries and the sample size calculation is visualised by calling `plot(design_RTSA)`. Figure 5 show a TSA plot generated from a `RTSA()` call with `type = "analysis"`.

3.2. Updating a prospective sequential meta-analysis

When updating a prospective sequential meta-analysis, we need the prespecified design and the accumulating data on the trials. The sequential boundaries are then updated to the data, as it might not be possible to have the exact allocation of participants and/or events in the trials as originally planned. As data is provided, a meta-analysis is calculated and conditional inference is also calculated. Three of the subfunctions are used when the `RTSA()` function is used for updating a design. As visualised on Figure 3, the `inference()` function depend on both the boundaries but also the results of the cumulative meta-analysis.

The argument `type = "analysis"` is used for specifying that the `RTSA()` function is used for analysing and not planning. The additional remaining arguments which can be used in the update of a prospective meta-analysis are:

```
RTSA(type = "analysis", data, design)
```

Now, we make toy-data to show how to analyse data with a prespecified design can be done using the `RTSA()` function and the design from subsection 3.1:

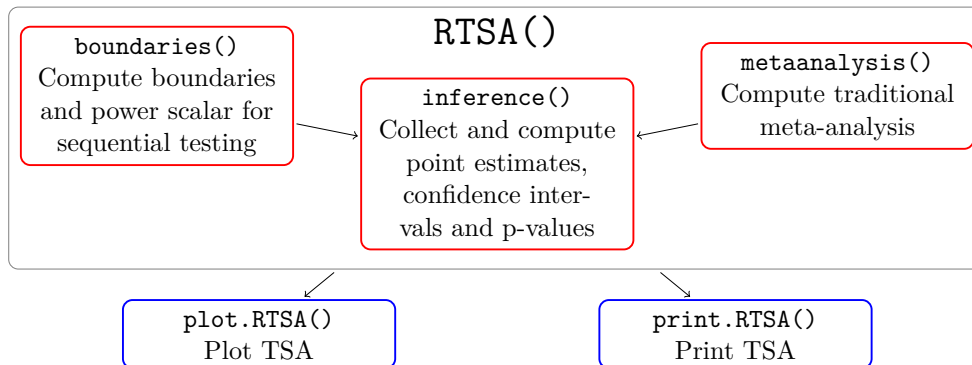


Figure 3: Subfunctions used in `RTSA()` and types of output when `RTSA()` is used for updating a prospective sequential meta-analysis with data.

```
R> data_example <- data.frame(study = c("A", "B", "C"), nI = c(1200, 1000, 500),
+                             nC = c(1100, 1100, 500),
+                             eI = c(70, 60, 60), eC = c(80, 80, 80))
```

Using the data we get:

```
R> RTSA(type = "analysis", data = data_example, design = design_RTSA)
```

Trial Sequential Analysis was computed with the following settings:

Boundaries for a 2-sided design with a type-I-error of 0.05, and type-II-error of 0.1.

Futility is set to: binding. Alpha-spending function: esOF.

Beta-spending function: esOF.

The required information size is not adjusted by heterogeneity. The required information size is further increased with 5 percent due to the sequential design. The total required information size is 9064.

Timing, boundaries, and test statistic:

sma_timing	upper	lower	fut_upper	fut_lower	z_fixed	z_random
0.263	4.333	-4.333	NA	NA	-0.844	-0.844
0.527	2.963	-2.963	0.299	-0.299	-1.811	-1.811
0.790	2.359	-2.359	1.251	-1.251	-2.478	-2.478
1.053	1.963	-1.963	1.963	-1.963	NA	NA

`sma_timing` is the ratio of the required sample for a sequential meta-analysis to a non-sequential meta-analysis sample size.

Timing, outcomes, and confidences intervals for fixed-effect and random-effects models:

```
sma_timing RR_fixed TSA_0.95lci_fixed TSA_0.95uci_fixed
```

0.263	0.875	0.441	1.736
0.527	0.813	0.579	1.141
0.790	0.792	0.634	0.989
1.053	NA	NA	NA
sma_timing	RR_random	TSA_0.95lci_random	TSA_0.95uci_random
0.263	0.875	0.441	1.736
0.527	0.813	0.579	1.141
0.790	0.792	0.634	0.989
1.053	NA	NA	NA

lci is the lower limit of the confidence interval. uci is the upper limit of the confidence interval.

Meta-analysis results:

Fixed pooled effect (RR): 0.79 (95% TSA-adjusted CI: 0.63;0.99), naive p-value: 0.0132

Median unbiased pooled effect (RR): 0.80 (95% SW-adjusted CI: 0.65; 0.96), SW p-value: 0.0173

Heterogeneity results:

tau²: 0.00; I²: 0.0%; D²: 0.0%; Heterogeneity p-value: 0.7358

Please note the following warnings:

- The order of the Trial Sequential Analysis will be based on the order of the studies in the data-set. Please add a 'order' column in the data-set to specify the order.

- Prob. of event in the control group is set to 0.1. The observed prob. of event is 0.0706. The power of the sequential might be affected.

The output from `RTSA()` describes that it was used for analysis and includes the boundaries for the sequential design. Note that the initial design does not change when updating the sequential meta-analysis with data. The timing and number of trials might be different than the original design which affects the stopping boundaries, but the required information size and whether the model is fixed or random does not change from the settings in the design. This means that if heterogeneity was not accounted for but is present, the sequential meta-analysis will be underpowered. If, however, the heterogeneity was smaller or not present when accounted for in the design the sequential meta-analysis will be overpowered. The power of the prospective sequential meta-analysis is also affected for larger deviations from the original number of trials and timings of the sequential meta-analysis. An estimate on how the power was affected by differences in the design compared to the actual observed data is part of future versions of the **RTSA** package.

3.3. Retrospective sequential meta-analysis

When there already is knowledge about trial results or some of the trials have been published that are going to be part of the meta-analysis, the meta-analysis becomes a retrospective sequential meta-analysis.

For an analyses of a retrospective sequential meta-analysis, `RTSA()` performs a retrospective sample size calculation for estimating RIS followed by an initial boundary calculation to calculate the scalar to provide the sequential adjusted estimate of RIS. The boundaries for the data provided are then calculated based on the observed data together with a naive meta-analysis. Conditional and unconditional inference are then calculated and collected. All the sub-functions are used when the meta-analysis is retrospective as shown in Figure 4.

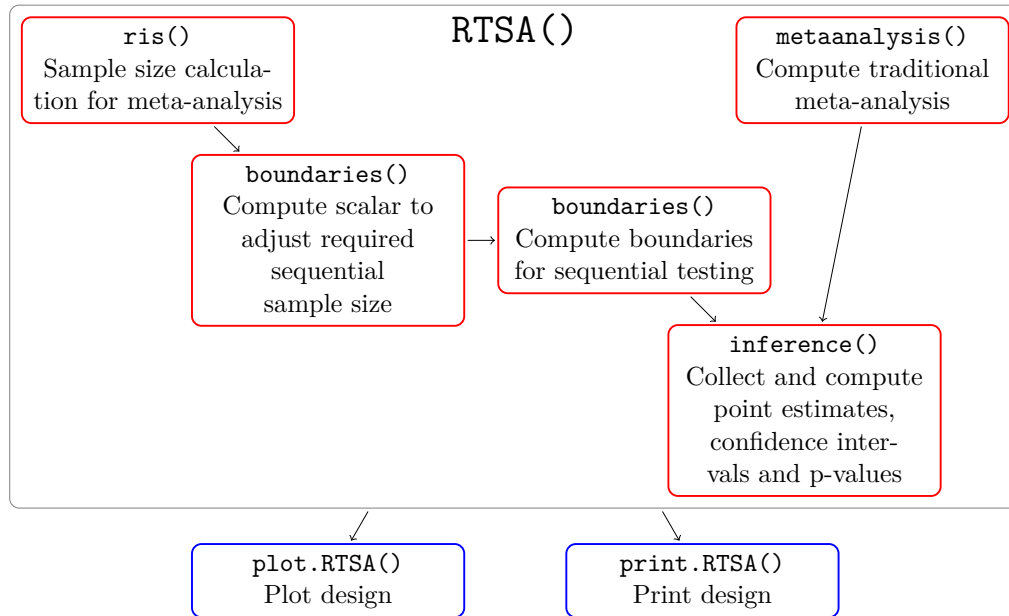


Figure 4: Subfunctions used in `RTSA()` and types of output when `RTSA()` is used for retrospective sequential meta-analysis.

To make a retrospective sample size calculation, we need to provide design parameters for the level of type-I-error, type-II-error, the minimum relevant clinical value, whether the design is two- or one-sided, the type of error-spending function, and if/how futility boundaries are used. Hence the design arguments are used when the meta-analysis is retrospective.

To illustrate the functionality, we use an existing meta-analysis as an example. The data example is concerned with the risk of infection during surgery. The hypothesis of interest is whether the fraction of oxygen provided during surgery affects the risk of surgical site infection during surgery. This null hypothesis has been investigated in multiple trials. A Cochrane review was performed on 15 trials, comparing an oxygen fraction of 60% to 90% versus a fraction of 30% to 40%. The primary outcome was surgical site infection [Wetterslev et al. \(2015\)](#). The perioperative inspiratory oxygen data can be found in the `RTSA` package and is named `perioOxy`:

```
R> data("perioOxy", package = "RTSA")
```

The trials in the `perioOxy` data set are shown in subsection 4.1 and in a forest plot in Figure 6. We will compute a retrospective sequential meta-analysis comparing trials with results on the difference in incidence of getting surgical site infection between low and high oxygen

fraction. We will assume that the meta-analysis was updated each time a new trial is added to the meta-analysis as per their publication year. Hence, we expect to have an interim analysis at each trial expect for trials which add less than 1% of the required information size. Using the `perioOxy` data from the **RTSA** package an example of a retrospective TSA is:

```
R> retro_RTSA <- RTSA(type = "analysis", outcome = "RR", pC = 0.129,
+                      mc = 0.8, side = 2, data = perioOxy, alpha = 0.05,
+                      es_alpha = "esOF", beta = 0.2, futility = "binding",
+                      es_beta = "esOF", re_method = "DL", random_adj = "D2")
```

The setup for the sequential meta-analysis is as follows. We defined the minimum relevant difference as 20% reduction in relative risk from the control group (RR of 0.8). Furthermore, we set the baseline risk of event to be the observed proportion of events relative to the sample in the control group, which is approximately 13%. The null hypothesis is no difference in relative risk between the different fractions of oxygen of the incidence of surgical site infection. Thus, a two-sided test is used. We set the significance level α to 5% and power to 80%. α - and β -spending functions will be Lan and DeMets' versions of O'Brien-Fleming's boundaries and the futility boundaries will be binding. The DerSimonian-Laird estimate will be used for estimation of the variance in the random-effects meta-analysis and the sample size will be adjusted by Diversity'.

As with the other two usages of the `RTSA()` function, the results of the retrospective TSA can be printed or plotted. Two types of plots are available when the `RTSA()` function is used for analysis. If interested in the test-statistic and the testing boundaries, `plot()` of the `RTSA()` call shows the boundaries and the observed test-statistic. If interested in the point estimate and confidence intervals `plot(..., type = "outcome")` can be called to visualize the process on the outcome scale such as, e.g., RR. Figure 5 shows the cumulative test statistic and the stopping boundaries as a function of the information fraction.

When the meta-analysis is retrospective and becomes sequential, there are several potential issues including sequential decision bias, sequential design bias, and reduced protection of the type-I-error [Kulinskaya et al. \(2015\)](#). This means that we cannot interpret point estimates, confidence intervals and p-values as valid. The retrospective meta-analysis is, however, still informative. It provides the best guess of the general treatment effect while setting the size of the current information available in context to the required information size. It further informs that a hypothesis cannot be tested sequentially without affecting the type-I-error. The exact level of control of type-I-error in a retrospective sequential meta-analysis is difficult to quantify. Using the **RTSA**, one can via simulations investigate different scenarios in context of the specific sequential meta-analysis to make statements about the expected type-I-error. How to use the package for simulation is not the focus of this paper but will be an essential part of a future article.

4. Subfunctions in RTSA

This section describes three subfunctions **RTSA** that are used in the `RTSA()` function, but can also be used independently.

4.1. `metaanalysis()`: Naive meta-analysis

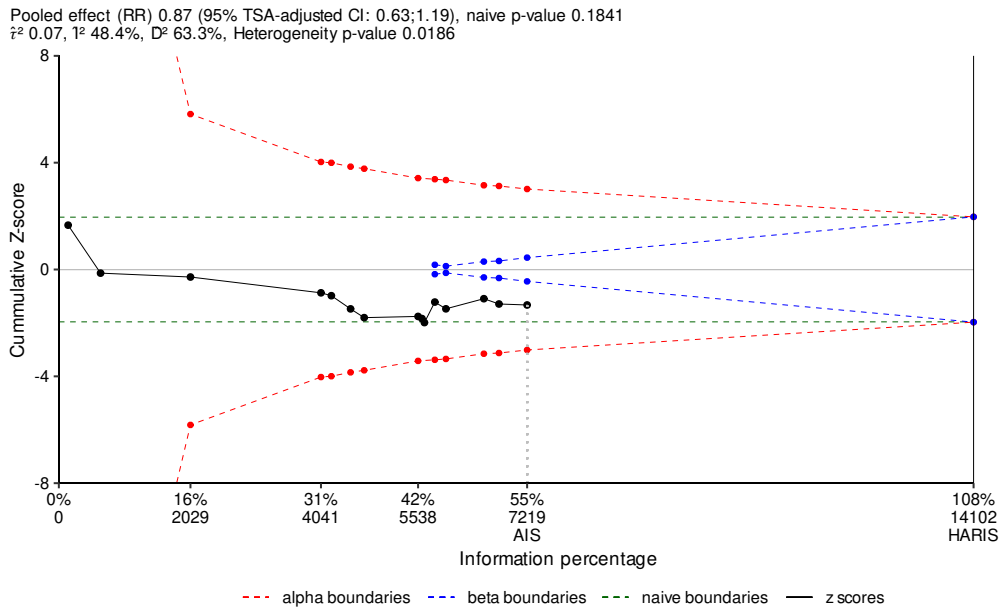


Figure 5: Plot generated for RTSA objects, here by calling `plot(retro_RTSA)`. Pictures the stopping boundaries and the cumulative test-statistic.

The function for meta-analysis in **RTSA** is `metaanalysis()`. The function computes a fixed-effect and a random-effects meta-analysis based on standard meta-analysis theory. To use the `metaanalysis()` function an outcome metric and the trials data must be included:

```
metaanalysis(outcome, data,...)
```

The function has several additional arguments to custom the specific meta-analysis. The arguments are either specific to the data input, the methods for conducting the meta-analysis, or to be used for a sample and trial size calculation. We list the main arguments in Table 3. Using the `perioOxy` data.frame provided in the **RTSA** package, we get the following output which shows the study and meta-analysis results.

```
R> metaanalysis(outcome = "RR", data = perioOxy)
```

Individual trial results:

	study	RR	se(log(RR))	lower.95CI	upper.95CI	w_fixed	w_random
1	Gardella 2008	1.823	0.601	0.897	3.704	3.12	5.67
2	Greif 2000	0.464	0.569	0.246	0.875	3.90	6.54
3	Meyhoff 2009	0.951	0.330	0.768	1.177	34.44	14.41
4	Myles 2007	0.740	0.378	0.559	0.979	19.96	12.96
5	Williams 2013	0.898	0.631	0.412	1.959	2.58	4.97

Argument	Description
<i>Outcome and data</i>	
<code>outcome</code>	Outcome metric. Options are "MD" (mean difference), "RR" (risk ratio), "OR" (odds ratio) and "RD" (risk difference).
<code>data</code>	a data.frame containing the set of columns <code>eI</code> , <code>eC</code> , <code>nI</code> , <code>nC</code> for binary data or the set of columns <code>mI</code> , <code>mC</code> , <code>sdI</code> , <code>sdC</code> , <code>nI</code> , <code>nC</code> for continuous data. The columns are respectively: <code>eI</code> , <code>eC</code> , <code>nI</code> , <code>nC</code> : <code>eI</code> is the number of events for the intervention group, <code>eC</code> is the number of events for the control group, <code>nI</code> is the number of participants in the intervention group and <code>nC</code> is the number of participants in the control group. <code>mI</code> , <code>mC</code> , <code>sdI</code> , <code>sdC</code> , <code>nI</code> , <code>nC</code> : <code>mI</code> is the mean effect in the intervention group, <code>mC</code> is the mean effect in the control group, <code>sdI</code> is the standard deviation of the effect in the intervention group, <code>sdC</code> is the standard deviation of the effect in the control group, <code>nI</code> is the number of participants in the intervention group and <code>nC</code> is the number of participants in the control group.
<i>Meta-analysis arguments</i>	
<code>weights</code>	Method for calculating weights. Options include "MH" (Mantel-Haenzel) and "IV" (inverse variance).
<code>re_method</code>	Method to estimate the variance of the random-effects model. Options include DerSimonian-Laird "DL" and the Hartung-Knapp-Sidik-Jonkman adjustment to DerSimonian-Laird "DL_HKSJ". Defaults to "DL_HKSJ".

Table 3: Main arguments used in `metaanalysis()`

6	Belda	2005	0.637	0.496	0.393	1.032	6.74	8.83
7	Bickel	2011	0.413	0.684	0.165	1.032	1.86	3.91
8	Duggal	2013	0.998	0.482	0.633	1.573	7.55	9.31
9	Golfam	2011	0.333	1.270	0.014	7.870	0.16	0.42
10	Mayzler	2005	0.667	0.747	0.223	1.990	1.31	2.95
11	Pryor	2004	2.222	0.607	1.078	4.580	3.00	5.52
12	Schietroma	2013	0.449	0.719	0.163	1.238	1.52	3.34
13	Scifres	2011	1.388	0.495	0.858	2.245	6.77	8.86
14	Stall	2013	0.706	0.572	0.372	1.340	3.82	6.45
15	Thibon	2012	0.920	0.594	0.461	1.836	3.29	5.87

Non-sequential metaanalysis results:

	type	RR	se(log(RR))	lower.95CI	upper.95CI	pValue
1	Fixed	0.880	0.064	0.777	0.997	0.0455
2	Random	0.869	0.105	0.707	1.069	0.1841

There are several elements that can be extracted from the meta-analysis object by calling `..$....`. The most important are:

- `..$study_results`: Individual study results
- `..$meta_results`: Meta-analysis results
- `..$hete_results`: Heterogeneity estimates and statistics
- `..$ris`: Retrospective sample and trial size calculation

The information contained by calling `study_results` and `meta_results` are almost identical to the information from the print. Results about heterogeneity is provided in the object and is printed when the meta-analysis is plotted by a forest-plot. The information can also be extracted by storing the meta-analysis and calling `..$hete_results`:

```
R> ma <- metaanalysis(outcome = "RR", data = perioOxy)
R> ma$hete_results

$hete_est
      Q Q_df    Q_pval      tau2      I.2      D.2
1 27.1138  14 0.01860805 0.06526157 0.4836577 0.6329575

$CI_heterogen

      estimate ci.lb  ci.ub
tau^2    0.0653 0.0026 0.3473
tau      0.2555 0.0510 0.5894
I^2(%)  48.3658 3.5997 83.2930
H^2      1.9367 1.0373 5.9855
```

Here the last table of information is made using the function `rma.uni()` from the **metafor** package Viechtbauer (2010). The purpose of the **RTSA** package is not to provide an extensive library of methods for traditional meta-analysis. We plan to extend the package with the possibility to add classes from the **metafor** package to allow for more custom meta-analysis methods to be used in the `RTSA()` function.

In **RTSA**, a forest plot can be used to visualise the trial and meta-analysis results. Such a plot can be created by calling `plot()` to the meta-analysis object, see Figure 6.

```
R> plot(ma)
```

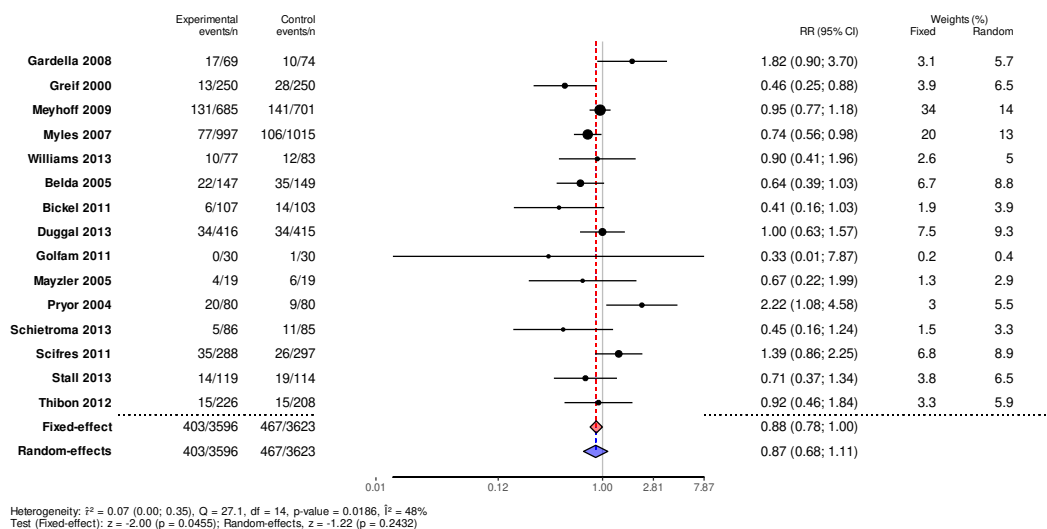


Figure 6: Forest plot from the **RTSA** package

If `mc`, which is the minimal clinical relevant value, is provided to the `metaanalysis()` function, a retrospective sample and trial size are calculated. By calling `..$ris` will a print be provided describing how many more participants are required to achieve the specified power. The default power is set to $1 - \beta$ which is 90% and if heterogeneity can be estimated a minimum number of extra trials are also provided. The following subsection describes the sample and trial size calculation in more detail.

4.2. `ris()`: Sample and trial size calculation

The **RTSA** package offers both a sample size and a trial size calculation for meta-analysis using the function `ris()`, *required information size*, providing the user with the required number of participants to reach a specific level of power when the meta-analysis is not sequential. The arguments of the function follow the notation from Section 2.1.1:

```
ris(outcome, mc, ...)
```

The `ris()` function can be used prospectively or retrospectively. Suppose we want to design a prospective meta-analysis with an expected RR of 0.8, a probability of 30% of event in the

control group, a type-I-error of 5%, type-II-error of 10% and an expected heterogeneity τ^2 of 0.01. Estimates of the sample and trial size for a fixed-effect and random-effects meta-analysis using the `ris()` function:

```
R> ris(outcome = "RR", alpha = 0.05, beta = 0.2, fixed = FALSE,
+      side = 2, mc = 0.8, pC = 0.3, tau2 = 0.01)
```

This is a prospective meta-analysis sample size calculation.

The sample size calculation assumes a 2-sided test, equal group sizes, a type-I-error of 0.05 and a type-II-error of 0.2.

The minimum clinical relevant value is set to: 0.8 for outcome metric RR.

Additional parameters for sample size are:

Probability of event in the control group: 0.3.

Fixed-effect required information size:

1720 participants in total.

Random-effects required information size:

Adjusted by τ^2 : 8186 participants in total split over (at minimum) 2 trial(s).

For more information about the sample size calculation see vignette:

'Calculating required sample size and required number of trials'.

If we assume that six centers are willing to participate in the prospective meta-analysis, one can add the number as an argument in the `ris()` function by setting `trials = 6`. The number of participants per trial is then provided along with the total number of participants by extracting `..NR_taunPax`.

```
R> ris_6 <- ris(outcome = "RR", alpha = 0.05, beta = 0.2, fixed = FALSE,
+              side = 2, mc = 0.8, pC = 0.3, tau2 = 0.01, trials = 6)
R> ris_6$NR_tau$nPax
```

	[,1]	[,2]	[,3]	[,4]	[,5]
Trials	2	3	4	5	6
Pax per trial	4093	1218	716	507	392
Total nr of pax	8186	3654	2864	2535	2352

The easiest way to make a retrospective sample and trial size calculation is to use the `metaanalysis()` function. A call to the `ris()` function is done in `metaanalysis()` function when the argument `mc` is provided. As some data is available, it is assumed that we are conducting a retrospective meta-analysis. If the required power is not reached, the output will inform about the additional number of participants required to have a well-powered meta-analysis.

The `ris()` function is called in `RTSA()` and can be extracted from any `RTSA` object by `..$ris`. Note that when a sequential design is imposed more participants are required compared to a non-sequential meta-analysis to achieve a specific power for the sequential meta-analysis. How many more are provided in the print of an `RTSA` object.

4.3. `boundaries()`: Group sequential boundaries

The stopping boundaries provided in `RTSA()` can be calculated separately using the `boundaries()` subfunction. The boundaries in **RTSA** are computed using theory on group sequential methods. The methods are based on [Jennison and Turnbull \(1999\)](#). To compute boundaries using **RTSA**, the following argument must be passed to the subfunction:

```
boundaries(timing, ...)
```

Here `timing` is the expected cumulative size of the trials relative to the required sample size for a non-sequential meta-analysis. Hence it is a vector of the amount of information achieved across the different meta-analyses, which is a proportion of the RIS from 0 to 1.0. Arguments `es_alpha` (α -spending) and `es_beta` (β -spending) can be used to specify the choice of error-spending functions. Options are "esOF", "esPoc", "HSDC" or "rho":

- "esOF" which calls `esOF(alpha, timing)`, the Lan and DeMets' version of O'Brien-Fleming spending function only dependent on the α and timing,
- "esPoc" which calls `esPoc(alpha, timing)`, the Lan and DeMets' version of Pocock spending function,
- "HSDC" which calls `HSDC(alpha, timing, gamma)`, the Hwang Sihi DeCani's error spending function, and,
- "rho" which calls `rho(alpha, timing, rho)`, the ρ -family spending function.

Additional arguments `gamma` and `rho` are used for respectively for the Hwang Sihi DeCani's error spending function and the ρ family error spending function.

The `boundaries()` function can be used to investigate what is optimal for one's specific scenario. The choices of the design such as timings, error spending functions, the choice of futility boundaries and more affect the probability to reject the null hypothesis at the interim analyses as well as the required sample and trial size for a sequential meta-analysis. The boundaries can be both printed and plotted using `print()` or `plot()`, respectively.

While **RTSA** in principle can be used for designing clinical trials, we strongly recommend to use packages specifically designed for these purposes such as [gsDesign Anderson \(2022\)](#) or [rpact Wassmer and Pahlke \(2022\)](#). Both of these packages are tailored to clinical trials.

5. TSA in R versus Java

We wanted to update the Trial Sequential Analysis (TSA) software with newer methods and it has been requested from users of the Java implementation to be able to do TSA in R. The **RTSA** package was originally planned to be a one-to-one translation of the Java implementation of TSA with some extra functionalities. However, with a thorough walk-through of the implementation several of the core elements of TSA ended up being modified. Below is a list of the most important changes and extensions of the original Java software [Thorlund *et al.* \(2011\)](#).

5.1. Change of numerical integration method

To calculate the stopping boundaries, numerical integration is used. The method originally implemented relied on the trapezoidal rule for integration but was changed to a quadrature rule with Simpson's rule for grid points and weights. The new numerical integration method follows chapter 19 in [Jennison and Turnbull \(1999\)](#).

5.2. Futility boundaries

The original implementation only allowed for two-sided non-binding futility boundaries which was not clear for the user [Thorlund *et al.* \(2011\)](#). Now both binding and non-binding futility boundaries are possible for both one-sided and two-sided tests.

5.3. Sample size calculation under the random-effects model

RTSA includes new methods to calculate the sample size for a random-effects meta-analysis based on the paper by [Kulinskaya and Wood \(2013\)](#) and derivations as shown in Subsection 2.1.1. Users of the package can now get the a sample size and the trial size estimate for both prospective and retrospective meta-analyses, meaning that the user can be provided with a trial and sample size prior to the meta-analysis and during the sequential meta-analysis.

5.4. Vignettes

The vignettes are thought as one of the main new contributions to the software which provides help for the users of **RTSA** in the design and analyses of sequential meta-analysis. The current vignettes are: Calculating required sample size and required number of trials, Standard operating procedure for RTSA, Futility, and Prospective and retrospective sequential meta-analysis.

5.5. Miscellaneous

Below is a list of other important developments and contributions:

- *Boundary calculation is dependent on whether the functions are used for a prospective design or for prospective or retrospective analysis.* When using a specified design in the prospective meta-analysis, the boundaries when analysing the sequential meta-analysis depend on the initial design, and the observed data. This is different to the retrospective meta-analysis which are almost entirely data-driven. This distinction was not made in the **Java** implementation. See the prospective and retrospective sequential meta-analysis for more information.
- *More error-spending functions to select from.* These are Lan and DeMets' versions of Pocock's boundaries [DeMets and Lan \(1994\)](#), the Hwang-Shih-DeCani's error spending function [?](#), and the power family [Wang and Tsiatis \(1987\)](#), [Emerson and Fleming \(1989\)](#), and [Pampallona and Tsiatis \(1994\)](#)
- *Core methods written in C++ to increase speed.*
- *Sample size inflation factor to achieve correct power.* A cost of using a sequential testing scheme is a cost in power. To achieve the wanted level of power, a scalar/inflation factor

is used to correct the loss in power [Jennison and Turnbull \(1999\)](#). The size of the factor depends on the timing on the trials used in the meta-analysis but will most often be less than 20%.

- *The Hartung-Knapp-Sidik-Jonkman adjustment to the DerSimonian-Laird estimator for heterogeneity in meta-analysis.* The Hartung-Knapp-Sidik-Jonkman adjustment is recommended when the number of trials in a meta-analysis is small [IntHout et al. \(2014\)](#).
- *Stage-wise-ordering for the inference after a sequential meta-analysis.* The stage-wise ordering is used for adjusting the estimates (point estimates, confidence intervals and p-values) after a sequential analysis has crossed a stopping boundary or reached RIS.

The above listed additions are the most important ones; however, the full code has been rewritten. Worth noting is that the **RTSA** generates comparable/identical results as compared to the Java implementation, if the settings are the same.

6. Discussion and future work

RTSA translates and updates the Trial Sequential Analysis software from Java to R. During the implementation several key elements was updated as specified in Section 5. The implementation to R allows for a greater transparency to the methods used in TSA. Furthermore, the vignettes and examples in the **RTSA** package support the intended usage of TSA to the users.

There are several points for future additions to **RTSA**. There exists many meta-analysis packages in R such as **metafor** [Viechtbauer \(2010\)](#), **meta** [Schwarzer et al. \(2015\)](#) and **dmetar** [Harrer et al. \(2019\)](#) which provide much more sophisticated methods for meta-analyses than **RTSA**. A natural extension to **RTSA** would be the ability to incorporate the meta-analyses from one or more of these packages. For complete transparency of implementation, we decided to implement our own methods for meta-analysis and for most usages the **RTSA** tools for meta-analysis should be sufficient.

The control of type-I-error is not clear-cut in retrospective sequential meta-analyses. As retrospective meta-analysis is implemented in **RTSA**, the users can use the package for simulation studies to investigate the control in different scenarios. A future vignette and article is planned to showcase the properties of retrospective sequential meta-analysis.

There is less focus on power calculations in meta-analyses than in single trials. As sample and trial sizes are used in TSA to design testing schemes, the meta-analysis will always have an accompanied sample and trial size calculation. The exact power of the current meta-analysis is, however, not presented to the user. Current power is planned to be part of future updates of the **RTSA** package.

The **RTSA** package is currently focused on testing the null hypothesis. Using futility boundaries adds another dimension to the testing scheme with allowing the user to stop the meta-analysis, when the risk of falsely rejecting the null hypothesis is low. However, none of the boundaries are directly making statements about the minimal clinical relevant value. We plan to incorporate another testing process to **RTSA** which will be directly related to the minimal clinical relevant value.

References

- Anderson K (2022). *gsDesign: Group Sequential Design*. R package version 3.3.0, URL <https://CRAN.R-project.org/package=gsDesign>.
- Armitage P (1957). “Restricted Sequential Procedures.” *Biometrika*, **44**(1-2), 9–26. doi: [10.1093/biomet/44.1-2.9](https://doi.org/10.1093/biomet/44.1-2.9). URL <http://dx.doi.org/10.1093/biomet/44.1-2.9>.
- Chalmers I, Bracken MB, Djulbegovic B, Garattinia S, Grant J, Gülmezoglu AM, Howells DW, Ioannidis JPA, Oliver S (2014). “How to increase value and reduce waste when research priorities are set.” *The Lancet*, **383**(9912), 156–165. ISSN 0140-6736. doi:[https://doi.org/10.1016/S0140-6736\(13\)62229-1](https://doi.org/10.1016/S0140-6736(13)62229-1). URL <https://www.sciencedirect.com/science/article/pii/S0140673613622291>.
- Deeks J, Higgins J, Altman D (2019). “Analysing data and undertaking meta-analyses.” In *Cochrane Handbook for Systematic Reviews of Interventions version 6.0*, chapter 10. Cochrane.
- DeMets DL, Lan KKG (1994). “Interim Analysis: the Alpha Spending Function Approach.” *Statistics in Medicine*, **13**(13-14), 1341–1352. doi:[10.1002/sim.4780131308](https://doi.org/10.1002/sim.4780131308). URL <https://doi.org/10.1002/sim.4780131308>.
- DerSimonian R, Laird N (1986). “Meta-Analysis in Clinical Trials.” *Controlled Clinical Trials*, **7**(3), 177–188. doi:[10.1016/0197-2456\(86\)90046-2](https://doi.org/10.1016/0197-2456(86)90046-2). URL [http://dx.doi.org/10.1016/0197-2456\(86\)90046-2](http://dx.doi.org/10.1016/0197-2456(86)90046-2).
- Emerson SS, Fleming TR (1989). “Symmetric Group Sequential Test Designs.” *Biometrics*, **45**(3), 905. doi:[10.2307/2531692](https://doi.org/10.2307/2531692). URL <http://dx.doi.org/10.2307/2531692>.
- Harrer M, Cuijpers P, Furukawa T, Ebert DD (2019). *dmetar: Companion R Package For The Guide 'Doing Meta-Analysis in R'*. R package version 0.0.9000, URL <http://dmetar.protectlab.org/>.
- Higgins JPT (2003). “Measuring Inconsistency in Meta-Analyses.” *BMJ*, **327**(7414), 557–560. doi:[10.1136/bmj.327.7414.557](https://doi.org/10.1136/bmj.327.7414.557). URL <http://dx.doi.org/10.1136/bmj.327.7414.557>.
- Higgins JPT, Thompson SG (2002). “Quantifying Heterogeneity in a Meta-Analysis.” *Statistics in Medicine*, **21**(11), 1539–1558. doi:[10.1002/sim.1186](https://doi.org/10.1002/sim.1186). URL <http://dx.doi.org/10.1002/sim.1186>.
- Hwang IK, Shih WJ, Cani JSD (1990). “Group Sequential Designs Using a Family of Type I Error Probability Spending Functions.” *Statistics in Medicine*, **9**(12), 1439–1445. doi: [10.1002/sim.4780091207](https://doi.org/10.1002/sim.4780091207). URL <http://dx.doi.org/10.1002/sim.4780091207>.
- Imberger G, Gluud C, Boylan J, Wetterslev J (2015). “Systematic Reviews of Anesthesiologic Interventions Reported As Statistically Significant.” *Anesthesia & Analgesia*, **121**(6), 1611–1622. doi:[10.1213/ane.0000000000000892](https://doi.org/10.1213/ane.0000000000000892). URL <http://dx.doi.org/10.1213/ANE.0000000000000892>.

- Imberger G, Thorlund K, Gluud C, Wetterslev J (2016). “False-Positive Findings in Cochrane Meta-Analyses With and Without Application of Trial Sequential Analysis: an Empirical Review.” *BMJ Open*, **6**(8), e011890. doi:10.1136/bmjopen-2016-011890. URL <http://dx.doi.org/10.1136/bmjopen-2016-011890>.
- IntHout J, Ioannidis JP, Borm GF (2014). “The Hartung-Knapp-Sidik-Jonkman Method for Random Effects Meta-Analysis Is Straightforward and Considerably Outperforms the Standard Dersimonian-Laird Method.” *BMC Medical Research Methodology*, **14**(1), 25. doi:10.1186/1471-2288-14-25. URL <https://doi.org/10.1186/1471-2288-14-25>.
- Ioannidis JPA (2022). “Correction: Why Most Published Research Findings Are False.” *PLOS Medicine*, **19**(8), e1004085. doi:10.1371/journal.pmed.1004085. URL <http://dx.doi.org/10.1371/journal.pmed.1004085>.
- Jennison C, Turnbull B (1999). *Group sequential tests with applications to clinical trials*. Chapman & Hall/CRC Interdisciplinary Statistics. Chapman & Hall, UK United Kingdom. ISBN 9780849303166.
- Jennison C, Turnbull BW (1984). “Repeated Confidence Intervals for Group Sequential Clinical Trials.” *Controlled Clinical Trials*, **5**(1), 33–45. doi:10.1016/0197-2456(84)90148-x. URL [http://dx.doi.org/10.1016/0197-2456\(84\)90148-x](http://dx.doi.org/10.1016/0197-2456(84)90148-x).
- Kulinskaya E, Huggins R, Dogo SH (2015). “Sequential Biases in Accumulating Evidence.” *Research Synthesis Methods*, **7**(3), 294–305. doi:10.1002/jrsm.1185. URL <https://doi.org/10.1002/jrsm.1185>.
- Kulinskaya E, Wood J (2013). “Trial Sequential Methods for Meta-Analysis.” *Research Synthesis Methods*, **5**(3), 212–220. doi:10.1002/jrsm.1104. URL <https://doi.org/10.1002/jrsm.1104>.
- Pampallona S, Tsiatis AA (1994). “Group Sequential Designs for One-Sided and Two-Sided Hypothesis Testing With Provision for Early Stopping in Favor of the Null Hypothesis.” *Journal of Statistical Planning and Inference*, **42**(1-2), 19–35. doi:10.1016/0378-3758(94)90187-2. URL [http://dx.doi.org/10.1016/0378-3758\(94\)90187-2](http://dx.doi.org/10.1016/0378-3758(94)90187-2).
- Pogue JM, Yusuf S (1997). “Cumulating evidence from randomized trials: Utilizing sequential monitoring boundaries for cumulative meta-analysis.” *Controlled Clinical Trials*, **18**(6), 580–593. ISSN 0197-2456. doi:[https://doi.org/10.1016/S0197-2456\(97\)00051-2](https://doi.org/10.1016/S0197-2456(97)00051-2). Eighth International Symposium on Long-Term Clinical Trials, URL <https://www.sciencedirect.com/science/article/pii/S0197245697000512>.
- Schwarzer G, Carpenter JR, Rücker G (2015). *Meta-Analysis with R*. Springer International Publishing Switzerland.
- Seidler AL, Hunter KE, Cheyne S, Gherzi D, Berlin JA, Askie L (2019). “A Guide To Prospective Meta-Analysis.” *BMJ*, **nil**(nil), l5342. doi:10.1136/bmj.l5342. URL <http://dx.doi.org/10.1136/bmj.l5342>.
- Simmonds M, Salanti G, McKenzie J (2017). “Living Systematic Reviews: 3. Statistical Methods for Updating Meta-Analyses.” *Journal of Clinical Epidemiology*, **91**(nil), 38–46. doi:10.1016/j.jclinepi.2017.08.008. URL <http://dx.doi.org/10.1016/j.jclinepi.2017.08.008>.

- Spence GT, Steinsaltz D, Fanshawe TR (2016). “A Bayesian Approach To Sequential Meta-analysis.” *Statistics in Medicine*, **35**(29), 5356–5375. doi:10.1002/sim.7052. URL <https://doi.org/10.1002/sim.7052>.
- Thorlund K, Engstrøm J, Wetterslev J, Brok J, Imberger G, Gluud C (2011). *User manual for trial sequential analysis (TSA)*. Copenhagen Trial Unit, Centre for Clinical Intervention Research, Copenhagen, Denmark. URL https://ctu.dk/wp-content/uploads/2021/03/2017-10-10-TSA-Manual-ENG_ER.pdf.
- Turner RM, Jackson D, Wei Y, Thompson SG, Higgins JPT (2014). “Predictive Distributions for Between-Study Heterogeneity and Simple Methods for Their Application in Bayesian Meta-Analysis.” *Statistics in Medicine*, **34**(6), 984–998. doi:10.1002/sim.6381. URL <https://doi.org/10.1002/sim.6381>.
- Viechtbauer W (2010). “Conducting meta-analyses in R with the metafor package.” *Journal of Statistical Software*, **36**(3), 1–48. URL <https://www.jstatsoft.org/v36/i03/>.
- Wang SK, Tsiatis AA (1987). “Approximately Optimal One-Parameter Boundaries for Group Sequential Trials.” *Biometrics*, **43**(1), 193. doi:10.2307/2531959. URL <http://dx.doi.org/10.2307/2531959>.
- Wassmer G, Brannath W (2016). *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Springer Series in Pharmaceutical Statistics. Springer International Publishing. doi:10.1007/978-3-319-32562-0. URL <https://doi.org/10.1007/978-3-319-32562-0>.
- Wassmer G, Pahlke F (2022). *rpact: Confirmatory Adaptive Clinical Trial Design and Analysis*. R package version 3.3.1, URL <https://CRAN.R-project.org/package=rpact>.
- Wetterslev J, Jakobsen JC, Gluud C (2017). “Trial Sequential Analysis in Systematic Reviews With Meta-Analysis.” *BMC Medical Research Methodology*, **17**(1), 39. doi:10.1186/s12874-017-0315-7. URL <http://dx.doi.org/10.1186/s12874-017-0315-7>.
- Wetterslev J, Meyhoff CS, Jørgensen LN, Gluud C, Lindschou J, Rasmussen LS (2015). “The Effects of High Perioperative Inspiratory Oxygen Fraction for Adult Surgical Patients.” *Cochrane Database of Systematic Reviews*, **2016**(9), nil. doi:10.1002/14651858.cd008884.pub2. URL <http://dx.doi.org/10.1002/14651858.CD008884.pub2>.
- Wetterslev J, Thorlund K, Brok J, Gluud C (2008). “Trial Sequential Analysis May Establish When Firm Evidence Is Reached in Cumulative Meta-Analysis.” *Journal of Clinical Epidemiology*, **61**(1), 64–75. doi:10.1016/j.jclinepi.2007.03.013. URL <https://doi.org/10.1016/j.jclinepi.2007.03.013>.
- Wetterslev J, Thorlund K, Brok J, Gluud C (2009). “Estimating Required Information Size By Quantifying Diversity in Random-Effects Model Meta-Analyses.” *BMC Medical Research Methodology*, **9**(1), 86. doi:10.1186/1471-2288-9-86. URL <https://doi.org/10.1186/1471-2288-9-86>.
- Whitehead A (2002). *Meta-analysis of Controlled Clinical Trials*. 2002 John Wiley & Sons, Ltd.

Affiliation:

Anne Lyngholm Sørensen
Section of Biostatistics
Department of Public Health
University of Copenhagen
1353 Copenhagen, Denmark
E-mail: als@sund.ku.dk

Journal of Statistical Software

published by the Foundation for Open Access Statistics

MMMMMM YYYY, Volume VV, Issue II

[doi:10.18637/jss.v000.i00](https://doi.org/10.18637/jss.v000.i00)

<http://www.jstatsoft.org/>

<http://www.foastat.org/>

Submitted: yyyy-mm-dd

Accepted: yyyy-mm-dd

8.3 Manuscript III

Adjusting for conditional bias in an updated meta-analysis arising from a decision to update based on promising results

Anne Lyngholm Soerensen & Ian C. Marschner

Details: Planning to be submitted to *Statistics in Medicine* in 2023.

RESEARCH ARTICLE

Adjusting for conditional bias in an updated meta-analysis arising from a decision to update based on promising results

Anne L. Soerensen^{1,2} | Ian C. Marschner³

¹School of Mathematical and Physical Sciences, Macquarie University, Sydney, Australia

²Section of Biostatistics, Department of Public Health, University of Copenhagen, Denmark

³NHMRC Clinical Trials Centre, University of Sydney, Sydney, Australia

Correspondence

Corresponding author Anne L. Soerensen, Section of Biostatistics, Department of Public Health, Øster Farimagsgade 5 opg. B, Postboks 2099 1014 København K, Denmark.
Email: als@sund.ku.dk

Abstract

The results of published trials and meta-analyses are used to motivate and justify the conduct of new trials. If the earlier analyses show promising but insignificant findings, it can be a justification for creating similar new trials to support the test of the research hypothesis. One way to combine the new evidence with the old is via meta-analysis. However, if the decision to update a meta-analysis is based on earlier results being promising, the updated meta-analysis might be biased. We call this bias conditional bias due to decision making in meta-analysis. This bias affects the point estimate. It occurs as promising results pointing in favour of the research hypothesis are more likely to motivate new trials than less motivating earlier results. This phenomenon has been described by others. In this paper, we propose a new estimator used to adjust for the bias of the point estimate in an updated meta-analysis. The new estimator is a conditional estimator conditioning on the special sample path of the test statistic, here specifically when the new trials were only created and added to the existing evidence because of a previous promising result. The estimator is motivated by methods used for bias adjustment in group sequential trials. We found that the penalized conditional estimator corrects for the conditional sequential bias in most scenarios. We compare the bias and variance of the new estimator with the naive estimator in both an application and in simulation studies.

KEY WORDS

cumulative meta-analysis, sequential meta-analysis, conditional sequential bias, sequential decision bias, conditional estimation

1 | INTRODUCTION

A meta-analysis synthesizes the results of several studies into an aggregated treatment effect estimate. Because it can be used for synthesizing the already existing knowledge, it is useful for motivating new studies. In this paper we will focus on the common situation where a later study is conducted conditional on a meta-analysis of earlier studies being promising but not definitive and then added to the existing evidence.

Gathering information on prior evidence before reporting on results from a clinical trial is required by leading medical journals¹. This requirement is consistent with the CONSORT statement, which recommends a discussion that includes the results of earlier studies if not a systematic review or meta-analysis, when reporting the findings of a clinical trial². While summarizing existing information would reduce research waste³, a side-effect might be biased inference if new studies are conducted conditional on the results of an earlier trial or a meta-analysis of earlier studies. The reason for this bias is that promising earlier results are more likely to motivate new almost similar trials than less promising or de-motivating earlier results. When previous meta-analyses are selected to continue based on the point estimate, it creates a type of selection bias and this bias will be present in a updated meta-analysis including both the new and old studies. Ellis and Stewart⁴ argue that if previous evidence was statistically significant, it might be more likely to motivate the conduct of similar studies, than had the findings shown little or no promise of the intervention or treatment. Ter Schure and Grünwald⁵ uses the term “Gold Rush” to describe how new studies are motivated by initial findings of statistical significance in earlier studies.

When available information is used in making the decision to conduct a new trial, we are in a scenario of making a sequential decision as described by Kulinskaya et al.⁶. They find that an updated meta-analysis will be biased if the probability of conducting a new trial is correlated with the size of the treatment effect in earlier studies. While the new trial will not be biased, it is the combination of the evidence motivating the new trial and the new trial that may be biased. This bias is called sequential decision bias and affects the updated meta-analysis by leading to an upwards bias. In this paper, we will focus on the bias of the intervention effect estimate in an updated significant meta-analysis that has been based on previous promising but not significant analysis or meta-analyses. We call this conditional bias due to decision making. This is a common situation in medical research and in Section 4 and 7 we provide some examples.

Updating the analysis or meta-analysis with a new study makes it sequential. We use the word sequential in this context without a formal sequential testing regime such as the ones in Trial Sequential Analysis⁷ or Sequential meta-analysis⁸ but simply to describe that evidence is accumulating. In addition to the potential upwards bias of the point estimate due to conditional bias due to decision making, other measures are affected by updating a meta-analysis. When the null hypothesis is repeatedly tested such as in a sequential meta-analysis, the type-I-error is affected. With careful planning the meta-analysis can be interpreted within a group sequential design (GSD) to control the type-I-error⁷. GSDs are originally used for randomized controlled trials (RCTs) with preplanned interim analyses⁹. In a trial using a GSD will the naively calculated point estimate be biased as a consequence of the design. Several methods has been proposed to adjust the bias of the point estimate such as conditional estimators which dependent on the time of stopping^{10 11 12}. Similarities between the conditional bias due to decision making in sequential or cumulative meta-analysis and the bias of the point estimate in group sequential trials are mentioned in previous works⁴ and⁶. Denne et al.¹³, present a bias adjustment of the point estimate for clinical trials that can be extended by more participants. This is similar to our scenario, where we update a trial or meta-analysis with more studies instead for adding more participants to the same study. In this paper we will investigate if the bias adjustment methods for sequential trials can be adapted to sequential meta-analysis and extended to adjust for the conditional bias due to decision making.

The purpose of the current paper is to introduce a method for adjusting the intervention effect estimate from a sequential meta-analysis of RCTs by extending existing bias adjustment methods for sequential trials to sequential meta-analyses where we expect a conditional bias due to decision making. The method takes into account the decision making process for conducting new trials based on previous promising trials. The structure of the paper is as follows. The first section introduces some basic meta-analysis and group sequential design theory including some comments on how to combine the two fields. We will then present three different types of estimators for the updated meta-analyzed intervention effect estimate. Two of the estimators are motivated by estimators previously proposed in the context of group sequential RCTs. This section will clarify some of the assumptions for using the methods we propose. A case study is then presented to show how the methods perform on real data which is followed by a simulation study showing the behaviour of the estimators. The paper ends with a discussion of the results.

2 | SEQUENTIAL META-ANALYSIS

This section starts with a brief introduction of the fixed-effect meta-analysis model followed by a presentation of the structure of a group sequential design. The section ends with comments on how to combine meta-analyses and sequential designs.

2.1 | Meta-analysis

A meta-analysis point estimate is calculated as a weighted average of the included studies' point estimates. Suppose we have decided to use K studies in the meta-analysis, where we define $k = 1, \dots, K$ to be the identifier of study. Each study reports an estimate y_k and a standard error s_k . In meta-analysis theory is the inverse of the standard error squared called information. The information is often used as weights in the weighted average. Let $I_k = 1/s_k^2$ be the information of study k and $w_k = I_k$ be the weight of study k . The pooled point estimate and variance of the estimate are then calculated as:

$$\hat{\theta} = \sum_k \frac{y_k \cdot w_k}{\sum_k w_k}, \quad \text{and} \quad \text{var}(\hat{\theta}) = \frac{1}{\sum_k w_k}. \quad (1)$$

This is known as a fixed-effect meta-analysis calculated using inverse-variance weighting. Often will the studies included in a meta-analysis differ. This can be caused by the research being held at different types of research sites or that the populations in

the studies are to some degree different. This introduces some expected heterogeneity, which can be accounted for in a random-effects meta-analysis. We will in this paper only show results using fixed-effect meta-analyses. The methods used later do not depend on whether the meta-analysis is fixed or random, hence all methods can be used for random-effects meta-analysis as well.

As we will consider sequential meta-analyses in this paper, we can extend the definition from (1) to take into account the repeated nature. Let $m = 1, \dots, M$ be the identifier of analysis m . Here $m = 1$ denotes the first analysis which may be of a single study or the first meta-analysis, $m = 2$ denotes the updated analysis which includes the studies from $m = 1$ and newly added studies, and so forth. We assume that no meta-analysis can be continue forever, hence we let M be the final meta-analysis. Let k_m be the set of studies used in meta-analysis m , where we note that the set of studies from a previous meta-analysis is included in a later meta-analysis, $k_{m-1} \in k_m$. We can then rewrite (1) as:

$$\hat{\theta}^{(m)} = \sum_{k_m} \frac{y_k \cdot w_k^{(m)}}{\sum_k w_k^{(m)}}, \quad \text{and} \quad \text{var}(\hat{\theta}^{(m)}) = \frac{1}{\sum_{k_m} w_k^{(m)}}. \quad (2)$$

At each meta-analysis in the series of sequential meta-analyses, a hypothesis test is conducted. Consider $\mathbf{Z}_m = (Z_1, Z_2, \dots, Z_m)$ to be the sequence of test statistics, one for each analysis up to current meta-analysis m . Here Z_1, Z_2, \dots, Z_m are standardised test statistics where $Z_m = \hat{\theta}^{(m)}/\text{var}(\hat{\theta}^{(m)})$ such as defined in ⁹.

We define a promising meta-analysis to be a meta-analysis with an observed test statistic pointing in the direction of interest but with an insignificant hypothesis test. Hence for a two-sided test with a significance level of 0.05, in a sequence of meta-analyses up to meta-analysis m , the $m - 1^{\text{th}}$ meta-analysis is defined to be promising when the following event occurs:

$$Z_{m-1}^{\text{promising}} = \{Z_{m-1} \in (-1.96, 0)\}. \quad (3)$$

This definition of promising will be used throughout this paper. However, the methods we present shortly can consider other regions of the test statistic than $(-1.96, 0)$.

Conducting an additional study conditional on (3) leads to a conditioning mechanism that may lead to upwards bias in the m^{th} meta-analysis by updating the previous meta-analysis with that additional study. Adding an unbiased trial to the existing evidence will to some degree moderate any bias stemming from updating the previous $Z_{m-1}^{\text{promising}}$ meta-analysis being promising. However, it will not remove it completely as one will continue using the the previous meta-analysis in the updated meta-analysis. Furthermore, the old evidence is not believed to be biased itself in this set-up, it is rather the decision to continue the analysis based on $Z_{m-1}^{\text{promising}}$ that makes the m^{th} meta-analysis biased. One can think of the old evidence being a sample from an unbiased experiment. All samples are however subject to variation and we consider those samples pointing in the direction of benefit to be more likely to be continued. If we believe the previous evidence to be of low quality or biased due to other reasons, the best solution might be to remove these trials. This is not the scenario that we are considering.

The goal of this paper is to find an unbiased estimator for estimating θ at meta-analysis m when the previous meta-analysis was promising and a motivation for conducting a new study. Thus, we want to condition on $Z_{m-1}^{\text{promising}}$. The theory of group sequential designs provides methods for adjusting intervention effect estimates conditional on a sequence of bounded test statistics, a set-up which fits naturally into the meta-analysis scenario just presented.

2.2 | Group sequential designs

Group sequential designs (GSDs) for trials protect the level of type-I-error under multiple testing of the same null hypothesis in a cumulative sample. The method allows for multiple looks during the progress of a trial, enabling testing for rejection of the null hypothesis at any of the looks. The looks during the trial are called interim analyses and the last analysis on the full sample is called the final analysis or end-of-trial analysis. If the null hypothesis is rejected at an interim analysis, the trial is allowed to stop at the interim analysis before reaching the full sample size. Consider a group sequential trial with K planned analyses with $K - 1$ interim analyses and a final analysis. The sequence of test statistics for this set-up is denoted $\mathbf{Z}_K = (Z_1, \dots, Z_K)$. Testing thresholds $\mathbf{b}_K = (b_1, \dots, b_K)$ are set to control for type-I-error in this set-up such that:

$$P\{|Z_k| \geq b_k \text{ for some } k = 1, \dots, K\} = \alpha.$$

Here α is the significance level. Given the stopping boundaries \mathbf{b}_K , the group sequential testing regime can be defined as follows. Suppose we have a two-sided symmetrical sequential test, we will then at interim analysis $k = 1$ to $k = K - 1$ continue the

analysis if the test statistic is between the testing thresholds $-b_k < Z_k < b_k$. If the test statistic at k is smaller or larger than the thresholds, $Z_k < -b_k$ or $b_k < Z_k$, the analysis stops for rejecting the null hypothesis. At the final analysis K if $Z_K < -b_K$ or $b_K < Z_K$ the meta-analysis rejects the null at the final analysis, but if $-b_K < Z_K < b_K$ the analysis failed rejection of the null and is stopped. An example of stopping boundaries b_K are shown on Figure 1 left plot for $K = 2$, where the Z -score is used as a testing statistic and the boundaries are Lan and DeMets' version of O'Brien-Fleming stopping boundaries¹⁴. The design is symmetric, such that the trial can stop for rejection of the null hypothesis in either direction with b_K being used as the upper stopping boundaries and $a_K = -b_K$ as the lower stopping boundaries.

The testing scheme affects the distribution of the estimated intervention effect. Take the first interim as an example, a point estimate will only be reported at the first interim if the boundary was crossed. This is different compared to one-analysis-only design where the point estimate will be reported at end-of-trial no matter whether the testing boundary was crossed. Consider the scenario of $\theta = \delta$ where $\delta > 0$. Given the random variation in the estimate of θ , regardless of the variation begin equally distributed to on each side of δ , it is most likely that the trial will stop at the first interim by crossing the upper boundary given $\delta > 0$. The point estimates at the interim will then overly represent the tail of the distribution of estimates for $\theta = \delta$. This results in an overestimated point estimate.

To reduce the bias from naive point estimate, the sample path of the test statistic can be used while taking the stopping time into account. At any later interim if the trial is stopped, the intervention effect point estimate is dependent on the fact that the trial did not stop at any previous interim analysis. Hence the sample space for the test statistics where the trial stops at interim k can be defined as:

$$\mathbf{Z}_k = \{Z_j \in (-b_j, b_j) \text{ for all } j \in 1, \dots, k-1, \text{ and } Z_k \notin (-b_k, b_k)\}, \quad (4)$$

where for $k = K$, $(-b_K, b_K) = \emptyset$. Several estimators condition on the sample space being defined as in (4) to get conditionally unbiased estimates of the intervention effect.

There is a close analogy between a group sequential trial and a sequential meta-analysis. Often will a goal of sequential meta-analysis be continued to either reach the required sample size or stop for significance, and if this is the case, the set up is the same as in a GSD trial but without the adjusted testing boundaries. Without the conditional bias due to decision-making the sample path of the test statistics for the updated meta-analysis might be expressed as:

$$\mathbf{Z}_m = \{Z_j \in (-1.96, 1.96) \text{ for all } j \in 1, \dots, m-1, \text{ and } Z_m \notin (-1.96, 1.96)\}, \quad (5)$$

With decision-making the process can instead be formulated as:

$$\mathbf{Z}_m = \{Z_j \in (-1.96, 1.96) \text{ for all } j \in 1, \dots, m-2, Z_{m-1} \in (-1.96, 0) \text{ and } Z_m \notin (-1.96, 1.96)\}, \quad (6)$$

Using that the sample space in (4) is very similar to the sample space in (6) for the set of promising meta-analyses, we can try to adapt the methods for GSD point estimates to the point estimates of a meta-analysis. We will use that there exist methods for conditioning on these types of sets.

2.3 | Sequential designs for meta-analysis

Group sequential designs adapted to sequential meta-analysis exist in various forms. These include, sequential meta-analysis by Whitehead¹⁵, sequential semi-Bayesian^{8,16} and sequential fully Bayesian meta-analysis¹⁷, and, Trial Sequential Analysis^{7,18}. However, the purpose of all these methods are to either describe or control for type-I-error, which is not the interest in this specific paper. We focus on the conditional bias. As discussed in Section 1, this type of bias is common in real meta-analysis contexts. To use a GSD in a sequential meta-analysis no matter for the purpose of type-I-error control or for our investigation of bias, one must have a definition of when the final analysis is reached. A sample size calculation can be used for this purpose. How to calculate the sample size for a meta-analysis depends on whether the meta-analysis is fixed or random. For methods to calculate the sample size see¹⁹ and²⁰.

It is possible that some sequential meta-analyses will follow a sequential testing scheme such as the one presented in Subsection 2.2. However, it is often the case that sequential meta-analyses continues to be tested naively. This situation is visualised on Figure 1 left plot, where naive testing boundaries is set for $k = \{1, 2\}$. As we are concerned with conditional bias due to decision making, the corresponding stopping boundaries/decision boundaries to this scenario is presented on the figure's middle plot. Here the analysis will only continue if the point estimate is pointing in the desired direction $\hat{\theta}^{(1)} < 0$ but the null

hypothesis could not be rejected $-1.96 < z_1 < 0$. Further, it is seen in the figure that we will assume that the first analysis did not adjust for multiple testing. Thus, we are considering 1.96 and -1.96 thresholds for whether the null hypothesis is rejected in the updated meta-analysis. Note that the boundaries becomes the naive boundaries after the decision to continue only promising analyses as we expect the test of the updated meta-analysis to be the standard two-sided test. The method we propose will be flexible towards other decision-making schemes such as sequential testing schemes as visualised in the right plot in Figure 1. For the sake of simplicity, we however stick to the naive testing boundaries for the majority of this paper. Thus we will be in the scenario visualised in the middle plot of Figure 1.

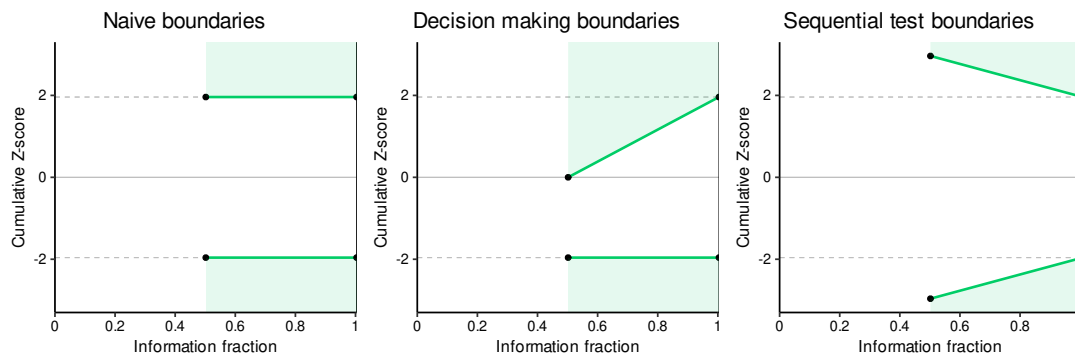


FIGURE 1 Stopping boundaries from respectively an updated meta-analysis with naive boundaries without conditional bias due to decision-making (left plot), an updated meta-analysis which may be biased from the conditional decision to continue the meta-analysis and a group sequential design (right plot). The green lines are the stopping boundaries and the coloured areas indicates where the null hypothesis will be rejected. The x-axis represents the information fraction. One way to define information fraction is by the accumulative sample compared to the total sample required.

In the group sequential trial setting, there has been created several estimators to handle the sampling path of the point estimate conditional on the stopping time, as the design of a sequential trial affects its distribution. Hence estimators have been created to handle the scenario considered in Figure 1 right plot. We will see whether we can extend these estimators to the scenario in the middle plot of the Figure and whether this adjustment will adjust for the conditional bias.

3 | ESTIMATORS

We consider three different point estimators in this paper. One is the naively calculated pooled point estimate and the other two are adjustments of the naive estimate. The first estimator considered is the naive estimator $\hat{\theta}^{(m)}$ from (2) at sequential meta-analysis m . The latter two estimators are adjusted versions of the naive estimator and are motivated by methods used in group sequential analysis of single studies. The first is the conditional estimator (CE)¹⁰ which conditions on the stopping time of the meta-analysis. Conditioning on the stopping time of the analysis leads to an adjustment that takes into account the distribution of estimates corresponding to that stopping time, which has the effect of removing bias. The second is the penalized conditional estimator (PCE)¹² which also conditions on the stopping time but uses regularization to deal with instability in the CE. A key novel feature of our application of conditional estimation to sequential meta-analysis is that we condition on earlier studies being promising but not statistically significant, which has the effect of adjusting for the sampling mechanism that introduces additional conditioning bias. We will now present the conditional estimators.

Conditional estimator (CE)

The size of the naive point estimate is dependent on which analysis the sequential meta-analysis is stopped at¹⁰. Both conditional and unconditional on the stopping time, the naive point estimate will be biased. A conditional unbiased estimate can be calculated by adjusting the naive estimate with a bias-correcting term. The bias can be expressed as follows. Let T be the stopping time, which was achieved at meta-analysis m where $m \in 1, \dots, M$. For $m < M$, the event $T = m$ is equivalent to the test statistic

crossing the stopping boundary for the first time at analysis m . For $m = M$, the event $T = m$ is equivalent to the test statistic not crossing the stopping boundary at analyses $1, \dots, M - 1$. With this specification of the event $T = m$, the conditional bias of the naive estimator is the difference between the true value θ and the expected value of the naive estimator at stopping time m ¹⁰:

$$B_{\theta}(\hat{\theta}^{(m)} | T = m) = \theta - E_{\theta}(\hat{\theta}^{(m)} | T = m). \quad (7)$$

Notice that the bias of $\hat{\theta}^{(m)}$ is a function of the true parameter value θ .

Undertaking a new study conditional on the earlier studies being promising is also a source of conditional bias in the naive estimator. In this situation, equation (7) needs to be modified to accommodate the additional condition, and the conditional bias becomes:

$$B_{\theta}(\hat{\theta}^{(m)} | T = m, Z_{m-1}^{promising}) = \theta - E_{\theta}(\hat{\theta}^{(m)} | T = m, Z_{m-1}^{promising}). \quad (8)$$

Here the bias is conditional on the time m at which the meta-analysis stopped and also that the previous $m - 1$ analysis showed a promising result. This last condition ($Z_{m-1}^{promising}$) is new and not included in previous works. Note in (8), that the expectation of the naive estimate is again a function of the true parameter value θ which is unknown. Our proposed approach will be to use this function to define an estimating equation that can be solved for θ and produces an estimate that is adjusted for the conditional bias. We now describe this approach.

Using the expression for the conditional bias in (8), we can define the adjusted estimators, which will be the naive estimator with the bias subtracted. We begin by defining an estimating equation by setting the conditional bias in (8) to zero:

$$B_{\theta}(\hat{\theta}^{(m)} | T = m, Z_{m-1}^{promising}) = 0. \quad (9)$$

By solving the estimating equation $B_{\theta}(\hat{\theta}^{(m)} | T = m, Z_{m-1}^{promising}) = 0$ for θ , we obtain an estimate that yields a conditional bias of zero, and is therefore adjusted for the conditional bias. This is the conditional estimator (CE) which we will write as $\bar{\theta}^{(m)}$ and which satisfies:

$$\bar{\theta}^{(m)} = \hat{\theta}^{(m)} - B_{\bar{\theta}^{(m)}}(\hat{\theta}^{(m)} | T = m, Z_{m-1}^{promising}).$$

Here $B_{\bar{\theta}^{(m)}}(\hat{\theta}^{(m)} | T = m, Z_{m-1}^{promising})$ is the bias-correction term for $\hat{\theta}^{(m)}$ evaluated at $\theta = \bar{\theta}^{(m)}$ and is dependent on the stopping time and the specific sample space for Z_1, \dots, Z_{m-1}, Z_m . To solve (9), we need an expression for $B_{\theta}(\hat{\theta}^{(m)} | T = m, Z_{m-1}^{promising})$. Per Fan et al¹⁰, we have that solving (9) is equivalent to maximizing the conditional log-likelihood of θ . Based on the conditional log-likelihood, the estimating equation is specified by:

$$B_{\theta}(\hat{\theta}^{(m)} | T = m, Z_{m-1}^{promising}) = \frac{1}{I_m} \cdot \frac{d}{d\theta} \log P_{\theta} (T = m, Z_{m-1}^{promising}) = 0. \quad (10)$$

Here I_m is the information at analysis m , hence $I_m = \text{var}(\hat{\theta}^{(m)})^{-1}$. Calculating the probability of stopping at analysis m given $Z_{m-1}^{promising}$, $P_{\theta} (T = m, Z_{m-1}^{promising})$ at θ , corresponds to calculating a multivariate normal integral. Different methods have been used to solve the almost similar problem of calculating $P_{\theta} (T = m)$, when adjusting for conditional bias in single group sequential trials. Liu et al¹¹ uses an iterative procedure (Newton-Raphson), whereas Marschner et al¹² provides an exact solution to the estimating equation by computation of the multivariate integral with R code provided in their Supplementary Material¹². We will re-purpose their code to our problem.

Although in principle the CE adjusts for the bias resulting from conditioning on $T = m$ and $Z_{m-1}^{promising}$, in practice there are situations in which this bias adjustment is unstable. This occurs in situations where a sequential meta-analysis crosses the stopping boundary but the test statistic is very close to the boundary. This phenomenon is well-known in sequential analysis of single studies and may produce an over-adjustment²¹. To correct this behavior, a penalized version of the conditional estimator can be used that regularizes this behavior. This is motivated by an approach used for sequential analysis of single studies where the stopping boundary is crossed¹². As we shall see in both the application and the simulation study, this is a useful approach when the test statistic is close to the stopping boundary.

Penalized conditional estimator (PCE)

To correct the over-adjusting behavior of the CE estimator a tuning parameter λ is introduced to control the over correction of the CE adjustment. The introduction of λ embeds the naive estimate and the CE within a class of estimates that allows the

choice of a favorable compromise between the two. The class of estimates is defined as:

$$\bar{\theta}_\lambda^{(m)} = \hat{\theta}^{(m)} - \lambda B_{\bar{\theta}_\lambda^{(m)}}(\hat{\theta}^{(m)} | T = m, Z_{m-1}^{promising}). \quad (11)$$

Here λ is a value between 0 and 1, where $\lambda = 0$ corresponds with the naive estimate $\bar{\theta}_\lambda^{(m)} = \hat{\theta}^{(m)}$ whereas $\lambda = 1$ corresponds to the CE, $\bar{\theta}_\lambda^{(m)} = \bar{\theta}^{(m)}$.

The key undesirable property of the CE is that it is possible for the test statistic to cross the stopping boundary and for the adjusted CE value to be on the opposite side of the null value. In other words, an analysis that concludes effectiveness may have an adjusted estimate in the direction of harm when using the CE. This is most likely to occur when the test statistic is close to the stopping boundary. By using equation (11), the tuning parameter λ can be used to regulate the adjustment so that this undesirable behavior cannot occur. We now describe how the value of λ is chosen. The value is chosen such that when the analysis is stopped in a given direction, such as $\hat{\theta}^{(m)} > 0$, the adjusted estimate will also point in the same direction $\bar{\theta}_\lambda^{(m)} > 0$. This can be achieved by choosing the value of λ that maps the adjusted penalized estimate at the stopping boundary onto 0. This is achieved by evaluating the bias at $\hat{\theta}^{(m)} = b_m / \sqrt{I_m}$, where b_m is the stopping boundary at analysis m . We can then calculate λ by setting $\bar{\theta}_\lambda^{(m)} = 0$ at the stopping boundary and isolate λ from (11):

$$\lambda = \frac{b_m / \sqrt{I_m}}{B_0(\hat{\theta}^{(m)} | T = m, Z_{m-1}^{promising})}. \quad (12)$$

This will ensure that whenever the test statistic goes beyond the stopping boundary, the adjusted estimate will be in the same direction. This is because the conditional estimates absolute value will increase as the absolute value of the naive estimate increases. As the penalized estimate can only be used for analyses that stopped early, we set $\lambda = 1$ when $m = M$. Hence the PCE is equal to the CE when the meta-analysis reaches the final analysis. In particular, we choose $\lambda = \lambda^*$ where,

$$\lambda^* = \begin{cases} \lambda \in [0, 1] : \text{As defined in (12)} & \text{for } m < M, \\ 1 & \text{for } m = M. \end{cases} \quad (13)$$

The adjusted estimate, which we call the penalized conditional estimate (PCE) is then defined using equation (11) as $\bar{\theta}_{\lambda^*}^{(m)}$. Similar to CE, the PCE is found by solving an estimating equation, determined by equation (11) in the same way as the CE estimating equation was determined. By introducing the penalty λ , the penalized estimator will be found in a two-step procedure where first the value of λ^* is determined before solving the estimating equation.

Both the CE and PCE were originally created for handling the conditional bias of stopping early in a single group sequential trial. We expect that extending the conditioning to also condition on $Z_{m-1}^{promising}$ and not just the stopping time will help in reducing the bias stemming from decision making in cumulative meta-analyses. Code for fitting CE and PCE has been written for the statistical software R²². The code is available in the online supplementary material in Marschner et al¹² which has been re-purposed to be used on meta-analyses rather than single trials.

Note that if we use the PCE and CE without the definition of promising, the estimators will be adjusting the point estimate of updated meta-analysis for stopping at the time it stopped. Stopping means that either the meta-analysis reached the required sample size with or without a rejection of the null hypothesis or it stopped early with a rejection of the null hypothesis. So the bias adjustment created by CE and PCE will adjust for two kinds of bias when also conditioning on the promising analyses being continued. It will adjust based on the promising analyses being continued but also the fact that it stopped. This means that it mostly makes sense to use the methods we will be proposing in when one uses the updated meta-analysis in a decision-making context. Thus, we are going to look at scenarios where the conductor of the updated meta-analysis will not recommend to continue the meta-analysis but consider the meta-analysis conclusive. This happens in the updated meta-analysis scenario once a certain sample size has been reached or the result is statistically significant.

4 | APPLICATION

To illustrate the application of our proposed approach in sequential meta-analysis, we will examine a published meta-analysis. The specific meta-analysis used as a case study investigates the effect of delayed cord-clamping versus early cord-clamping on in-hospital death of pre-term infants²³. The meta-analysis is conclusive with a p-value less than 0.05. Several trials have investigated whether there is an effect and a previous Cochrane review showed a promising effect of delayed cord clamping, but

the effect was not significant²⁴. Updating the meta-analysis with subsequent studies, including the large Australian Placental Transfusion Study (APTS)²⁵, showed that delayed clamping reduced in-hospital mortality significantly^{23,26}. The Cochrane review is mentioned in both the later meta-analysis and in a design paper describing the need for APTS²⁷. Both papers describe the promising results of the Cochrane review and a need for more evidence to be able to conclude whether in-hospital mortality is decreased by delayed cord-clamping. We will investigate whether the point estimate of the intervention effect when updating the Cochrane meta-analysis with APTS could potentially be biased as the justification of APTS was based on initial promising findings from the Cochrane review.

4.1 | Traditional meta-analysis

We present the naive updated analysis before investigating its potential bias. We consider first the original Cochrane meta-analysis which we then update with APTS. We can reproduce the point estimate from the original meta-analysis. The studies used and the results are found on the forest plot visualised in Figure 2. Here Cochrane refers to the original meta-analysis.

A motivation of the large APTS was the need for adequate power. The papers justifying APTS quote the Cochrane review's recommendation that future studies should include more data^{24,27}. We can calculate by how much the review was underpowered using the formula:

$$Power = 1 + \Phi(-Z_\alpha - \delta/\sqrt{V_F}) - \Phi(Z_\alpha - \delta/\sqrt{V_F}). \quad (14)$$

Here Φ is the CDF for the standard normal distribution, Z_α is the testing threshold (here 1.96), δ is the intervention effect of interest and V_F is the variance of the fixed-effect model²⁸. Thus, we need a value for δ and V_F . In the design of the APTS study the sample size was based on a relative reduction of approximately 25-30% (RR 0.7-0.75). In the Cochrane review the observed reduction was 37% (RR 0.63). Assuming RR values of 0.65, 0.7 and 0.75, we can use formula (14) to calculate the power for different values of δ . V_F is set to the variance of the point estimate observed in the Cochrane review. Table 1 shows the power calculated from the various settings.

TABLE 1 Power calculations based on the Cochrane review under varying values of the minimal clinically relevant relative risk reduction δ .

δ (RR)	0.63	0.7	0.75	0.8
Power	24%	16%	12%	9%

It is also of interest to know for which sample size, we would have a well-powered meta-analysis. The following formula can be used for this purpose^{7,19}:

$$RIS = 4 \cdot (Z_{1-\alpha/2} + Z_\beta)^2 \cdot \frac{\nu}{\theta^2}.$$

Here RIS stands for required information size, $\nu = (1 - p_A)p_A$, $p_A = (p_I + p_C)/2$ and $\theta = p_C - p_I$, with p_I being the probability of event in the intervention group and p_C being the probability of event in the control group. For the sample size calculation we will be assuming a range of control group mortality risks of 5%, 7.5% and 10%, a significance level of 5% and a power of 80%. Table 2 shows the sample size calculations.

Assuming that a minimum power of 80% is of interest for the researchers, looking at Table 1, we find that the Cochrane meta-analysis was under-powered for all considered values of δ with maximum power being 24%. This fits with the review's recommendation of creating new larger studies. Table 2 provides the sample sizes for which we will have 80% powered trials. We find from the table that adding APTS would achieve the adequate power for the updated meta-analysis under the scenario

TABLE 2 Sample size calculations based on different values of δ and p_C . Note that the number of participants in the Cochrane review was 458, thus the required sample size is not reached for any of the combinations of δ and p_C .

	$\delta = 0.63$	$\delta = 0.7$	$\delta = 0.75$	$\delta = 0.8$
$p_C = 0.05$	3586	5678	8406	13493
$p_C = 0.075$	2340	3702	5476	8783
$p_C = 0.1$	1717	2713	4011	6428

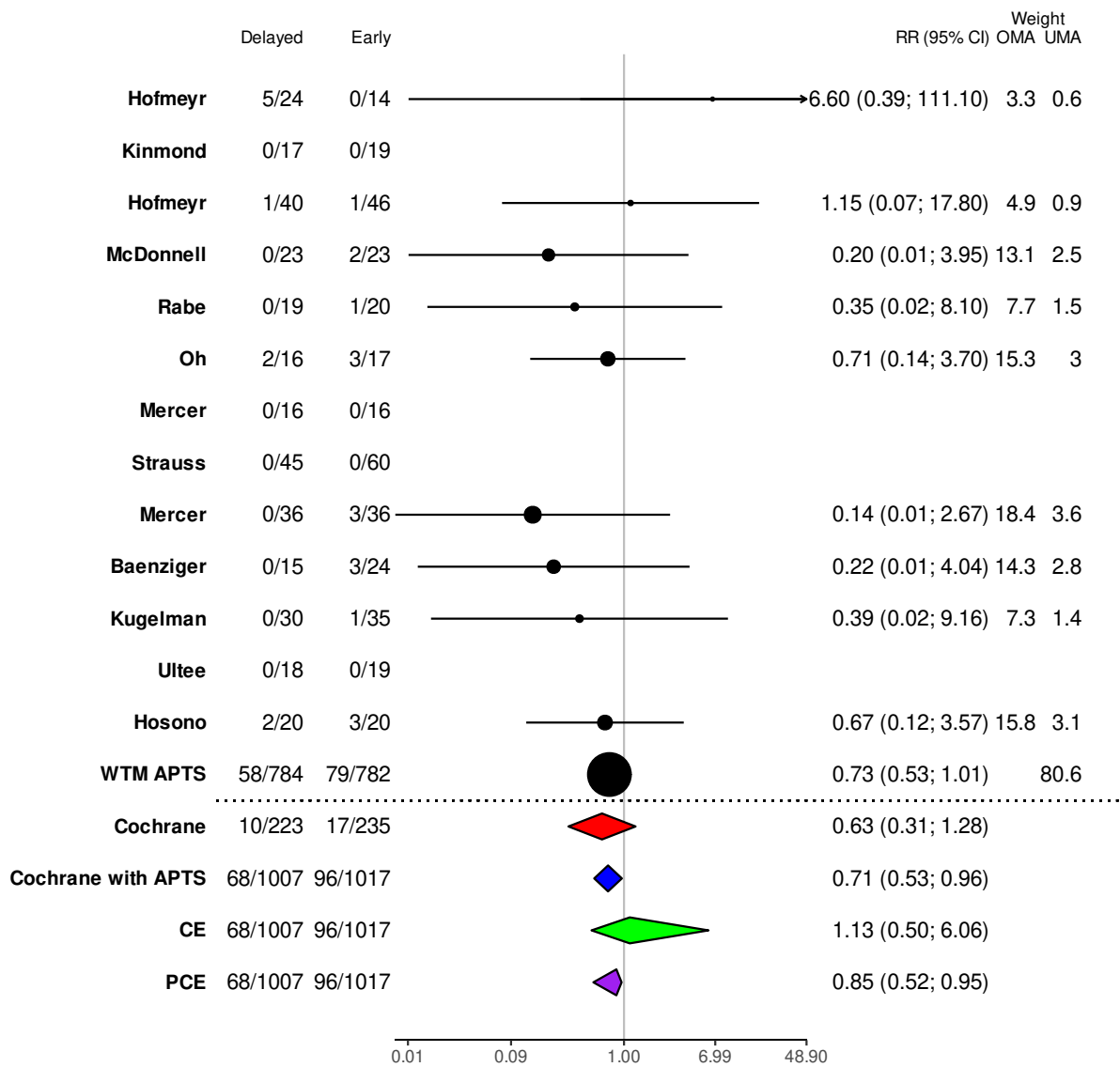


FIGURE 2 The original meta-analysis (OMA) and the updated meta-analysis (UMA). CE and PCE are the estimates of the conditional estimator and the penalized conditional estimator found in Table 6 with parametric bootstrap confidence intervals.

that $\delta = 0.63$ with $p_C = 0.1$ as the participant count in the Cochrane review is 458 and the participant count in APTS is 1566, which in total is 2024. As the design of APTS was created expecting a relative risk reduction between 25% and 30%, a larger sample size would have been needed to reach a power of 80%.

Adding APTS to the meta-analysis does result in a conclusive meta-analysis as seen from Figure 2 (Cochrane updated with APTS) and Table 3 with an estimate of the RR of 0.71 and an estimated $p_C \approx 10\%$. But the initiation of APTS and its study size was justified by looking at the results from the Cochrane systematic review. This makes the conduct of the study dependent on the results from the previous studies. We will investigate if updating the Cochrane review with APTS could potentially be biased and how to adjust for this bias using our proposed conditional estimators.

TABLE 3 Point estimates and z-values for the original meta-analysis and the updated meta-analysis.

	Point estimate (RR)	z-value	p-value
Cochrane	0.63	-1.28	0.1989
Cochrane updated with APTS	0.71	-2.26	0.0239

4.2 | Adjusted sequential meta-analysis

The updated meta-analysis is conclusive under the assumption of a two-sided test with a significance level of 0.05. The Cochrane meta-analysis also used a two-sided test with a significance level of 0.05. It is then possible that a sampling scheme as found in Table 4 is reflecting the decision making process of the trialists, when assuming that the second meta-analysis would not be made had the outcome of the first not been promising. This does not mean that the Cochrane meta-analysis tested with thresholds at 0 and -1.96, it means that it is assumed that the APTS was only started due to an earlier promising but not significant effect. As adding the participants from the APTS study to the Cochrane review will not reach the previously calculated RIS with an anticipated effect between 25% and 30%, we assume that more studies might be added if the update from APTS does not show significance.

TABLE 4 Adjusted naive sequential stopping boundaries.

	First meta-analysis	Second meta-analysis
Upper boundary	0	1.96
Lower boundary	-1.96	-1.96

Using these boundaries, we find that adding the APTS crosses the -1.96 boundary at the second meta-analysis with a z-value of -2.26. This is shown on Figure 3 left plot which further visualizes that we only consider promising meta-analyses by only allowing continuation if the z-score is between 0 and -1.96 for the first meta-analysis. We can in this scenario compute the MLE, CE and PCE. A sequential testing regime with the decision to only continue if the z-score is pointing in the direction of benefit but is in-significant is also shown on Figure 3. Here we see that the test statistic also cross the stopping boundary at the second analysis. The results of using CE and PCE in this scenario is presented in the appendix section A.

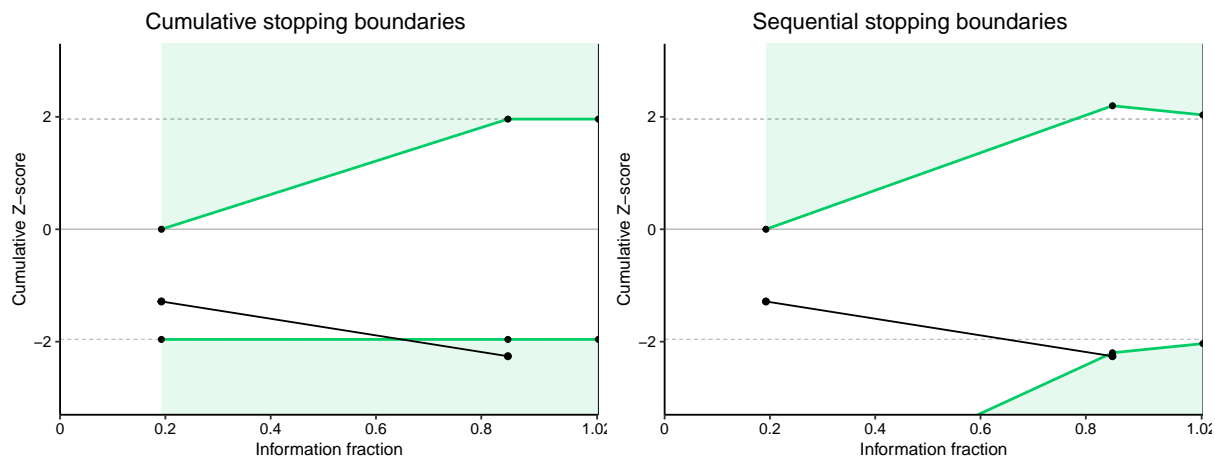


FIGURE 3 Stopping boundaries under a naively updated meta-analysis (left plot) and an updated meta-analysis with a sequential testing regime (right plot). In both plots are the assumption that only insignificant analyses pointing in the direction of benefit are continued. In both scenario is the updated meta-analysis reaching a significant result.

As described earlier, we compute the estimates by re-purposing code used for analysis of single clinical trials¹². The arguments used in the computation are shown in Table 5 where K is the number of expected analyses, k is the current analysis, stop side is defining that we stopped at the lower boundary, information is the inverse of the variance estimate at the first meta-analysis and second meta-analysis respectively and sided is whether the test is 1- or 2-sided.

TABLE 5 Arguments used to fit CE and PCE. We further use the stopping boundaries defined in Table 4.

	MLE	K	k	stop side	information	sided
Value	-0.3845	3	2	lower	7.50, 44.06	2

Using the arguments from Table 5, we get the estimates found in Table 6. To assess uncertainty in the estimates, we created 95% confidence intervals using both parametric and non-parametric bootstrap. The PCE is calculated with $\lambda^* = 0.8396$ by (12). We see that the CE crosses the null effect with a point estimate pointing in the opposite direction and both types of confidence intervals have limits in favour for both interventions. The PCE is also more conservative than the MLE, by being closer to the null, but does not cross the null effect in its confidence intervals. Hence the inference that the effect of delayed clamping is significant is retained when using the PCE. The PCE value indicates that the naive point estimate could be biased towards a greater effect due the estimate being closer to the null than the naive estimate.

TABLE 6 Point estimates (RR) with 95% confidence intervals.

	MLE (95% CI)	CE (95% CI)	PCE (95% CI)
95% CI - parametric bootstrap	0.7116 (0.5; 0.71)	1.1305 (0.5; 6.06)	0.8545 (0.5; 0.95)
95% CI - non-parametric bootstrap	0.7116 (0.52; 0.73)	1.1305 (0.52; 39.98)	0.8545 (0.52; 0.98)

The CE and PCE are affected by how close the statistic of the final analysis was to the boundary. Had we stopped closer to the stopping boundary, the CE will be pushed further in the opposite direction of the MLE. The Appendix section A contains an illustrative example of this using boundaries from a group sequential design on this case study. The estimates are also affected by the amount of information provided at the different analysis times, in the Appendix section B we are investigating the effect of having more information at the first meta-analysis.

We have now shown the use of the CE and PCE on a case study. The next section will investigate the bias and variance of the PCE, CE and MLE in a simulation study.

5 | SIMULATION STUDY

We wish to investigate the bias and variance of the different estimators. The scenario considered in this paper consists of an updated meta-analysis where the previous analysis (trial or meta-analysis) is promising. Thus the original analysis is only continued/updated with the addition of a new trial due to the original trial or meta-analysis being promising. We will in the simulation study replicate this mechanism. To simulate this as simple as possible, we start with simulating one trial which is then updated with a new trial in a meta-analysis, if the first trial's test statistic is promising. Trials that are not promising will not be continued. In reality they may be continued, but then we are no longer in the scenario that this paper is investigating. We are only considering the scenario where there might be a concern that the updated meta-analysis is affected by potential conditional bias due to decision-making. Here the specific decision is to only continue promising insignificant analyses.

A more elaborate technical explanation of the simulation set-up is explained below before we present the results.

5.1 | Simulation set-up

10,000 meta-analyses are simulated per scenario of interest. The meta-analyses are combining two studies sequentially where the outcome of interest is binary and the intervention effect is described using an odds ratio. The first analysis (only of the first study) must show promising results which we define to be a z -score between -1.96 and 0. A second study is then simulated and

combined with the first study in a meta-analysis. We will investigate both the scenario of having a conclusive meta-analysis before reaching the full sample size and reaching the full sample size. For simplicity there is no heterogeneity of the intervention effect in the simulation and a fixed-effect meta-analysis is used. We consider multiple scenarios expressed by a set of pre-specified parameters found in Table 7.

TABLE 7 Set of parameter values used in the simulation study. Notice that we define information fraction as the percentage of participants out of the required sample size.

Variable	Values
p_C - probability of event in control group	0.05, 0.2, 0.3, 0.5
θ (OR) - the true intervention effect	δ , 1, 1.1
δ (OR) - anticipated effect. Used for sample size calculation	0.3, 0.5, 0.7, 0.9
First information fraction (IF_1)	0.25, 0.5, 0.75
Second information fraction (IF_2) (full sample)	1, 1, 1
Second information fraction (IF_2) (conclusive before full sample)	0.75, 0.75, 0.9

The simulation scheme is as follows. The first three steps are prior to the 10,000 simulations.

1. Set the true odds ratio θ and p_C . p_C is the probability of event in the control group.
2. Calculate the full sample size N based on the assumed intervention effect δ and p_C . The sample size for the first study is the required sample size times the desired information fraction at the first analysis IF_1 (between 0.25 and 0.75). We consider equal sample sizes $N \cdot IF/2 = n_{1i}$ per treatment group i for $i \in \{1, 2\}$. The subscripts in n_{1i} stand for the first study sample size for the i^{th} treatment group.
3. Introduce the treatment difference using the two formulas:

$$\log\left(\frac{p_1}{1-p_1}\right) = \log\left(\frac{p_C}{1-p_C}\right) + \log(\delta).$$

$$\log\left(\frac{p_2}{1-p_2}\right) = \log\left(\frac{p_C}{1-p_C}\right).$$

Isolate p_1 and p_2 .

4. Simulate the number of events in treatment group 1 and 2, by use of the binomial distributions:

$$\hat{y}_{k1} \sim \text{Bin}(p_1, n_{k1}) \quad \text{and}$$

$$\hat{y}_{k2} \sim \text{Bin}(p_2, n_{k2}).$$

In case of total zero events, the simulation is redone. In the case of either treatment group having a 0 event count, 0.5 is added to each event count and 1 is added to each observation count. Note that total zero events and zero event counts happened rarely and did not influence the simulation results.

5. Calculate the z -score. If the z -score is not between -1.96 and 0, redo step 4. If the first analysis z -value is between -1.96 and 0, we run step 4 with $k = 2$.
 - (a) For meta-analyses that stop before reaching the full sample size, we have that $n_{2i} = n \cdot (IF_2 - IF_1)/2$ is the sample size for the treatment groups of the second trial.
 - (b) For meta-analyses that reach the full sample, we have that $n_{2i} = n \cdot (1 - IF_1)/2$ is the sample size for the treatment groups of the second trial.
6. Perform a meta-analysis of the two trials.
 - (a) In the scenario of meta-analyses stopping before reaching the full sample size, if the meta-analysis z -score is not below -1.96, re-run steps 4 to 6 until the meta-analysis z -score is below -1.96. Estimate and store MLE, CE and PCE.
 - (b) In the scenario of meta-analyses reaching the full sample size, estimate and store MLE, CE and PCE.
7. Continue steps 4 to 6 until 10,000 simulations are reached.

With the different values of p_C and δ from Table 7, we get the required sample sizes shown in Table 8.

We will present meta-analyses that end at the final analysis and meta-analyses that are conclusive prior to reaching the full sample size. For meta-analyses that reach the full sample size, we consider just the MLE and the CE as the PCE is equal to

	OR = 0.3	OR = 0.5	OR = 0.7	OR = 0.9
$p_C = 0.05$	186	813	3957	54932
$p_C = 0.2$	84	299	1299	16766
$p_C = 0.3$	82	260	1056	13010
$p_C = 0.5$	104	278	1004	11330

TABLE 8 Required sample sizes given p_C and OR

the CE in this scenario. Under the scenario of conclusive meta-analyses prior to reaching the full sample size the PCE and the CE differ and the performance of all three estimators will be investigated. For the PCE and the CE, we set the boundaries for determining the promising results to 0 and -1.96 at the first analysis. We will also investigate what happens if we do not condition on the first analysis being promising in the adjustment estimator. We call this estimator CE-. It is important to note that some of the scenarios we are investigating can have small probability of occurring. Table 9 show some of the probabilities:

TABLE 9 Estimated probability of either reaching the full sample size (IF2 = 1) or becoming conclusive earlier (IF2 < 1) at the second interim and having promising boundaries (-1.96 to 0) at the first interim.

	θ	δ	p_C	IF1	IF2	Probability
Reaching full sample size and $z_1 \in [-1.96, 0]$	0.7	0.7	0.3	0.25	1.0	64%
Reaching full sample size and $z_1 \in [-1.96, 0]$	1.0	0.9	0.05	0.5	1.0	46%
Reaching full sample size and $z_1 \in [-1.96, 0]$	1.1	0.9	0.2	0.75	1.0	6%
Conclusive at the second interim and $z_1 \in [-1.96, 0]$	0.7	0.7	0.3	0.25	0.75	23%
Conclusive at the second interim and $z_1 \in [-1.96, 0]$	1.0	0.9	0.05	0.5	0.75	1%
Conclusive at the second interim and $z_1 \in [-1.96, 0]$	1.1	0.9	0.2	0.75	0.9	<0.01%

Given being conclusive before reaching the full sample is so unlikely (probability less than 1%) when having a true value of $\delta = 1.1$, we will not report on these scenarios. Thus $\delta = 1.1$ will only be considered in the situations where we reach the full sample size. The results of Table 9 also highlight which of the scenarios are the most likely. For the scenarios of reaching the full sample size, the probability of reaching the full sample is highest when $\theta = \delta$ or $\theta = 1$. For the scenarios of being conclusive before reaching the sample size, the situation where $\theta = \delta$ is of highest probability.

In the simulations we use the same definition of promising as has been used throughout, see (3). It is possible to change this definition to $Z_{m-1} \in (-1.96, x)$, where x is some relevant lower threshold for continuing but this is not considered in the simulations presented here. As we are proposing the estimators to be used in scenarios where one might be cautious about the conditional bias due to decision-making, we will not consider introducing randomness regarding the decision-making definition into the simulations. While the choice of the lower threshold in the decision-making process, above defined as x , might be considered a random variable, for each use-case it will or should not be. The investigator should have some idea about when continuation of the analysis is justified which we assume to not be random. The purpose of the simulations is understand how the adjustment estimators behave in terms of bias and variance under the assumption of being used in practical examples.

5.2 | Results

Three general scenarios of the true intervention effect θ were considered as described in Table 7. One where $\theta = \delta$, hence the true effect was equal to the anticipated effect of intervention. One where there was no effect of intervention, thus $\theta = 1$ and a last scenario where $\theta = 1.1$.

5.2.1 | Reaching required sample size

In this section we will look at simulation results when the sample size is reached. In this scenario the CE and the PCE will be the same and we report the estimate as CE. We denote CE- to be the conditional estimator where we do not condition on the previous meta-analysis being promising. Table 10 is presented below which concerns the scenario of $\theta = \delta$ and $\theta = 1$ (no intervention effect). The last scenario for $\theta = 1.1$ (positive effect of control over intervention) is provided in the Additional results section, see Section C Table C4.

TABLE 10 Simulation results based on 10000 meta-analyses where the true intervention effect is $\theta = 1$ and $\theta = \delta$ and the full sample was reached. The first three columns are simulation settings. The remaining columns are the simulation results where bias is the average point estimate minus the true effect θ . IF_1 stands for the information fraction of the first study and translates to the fraction of participants in the first study compared to the full participant size. The event probability p_C is set to 0.2.

θ	δ	IF_1	Average point estimate			Bias			Standard deviation		
			MLE	CE-	CE	MLE	CE-	CE	MLE	CE-	CE
1	0.5	0.25	0.91	0.9	1.01	0.09	0.1	-0.01	0.27	0.29	0.33
1	0.5	0.50	0.87	0.85	1.03	0.13	0.15	-0.03	0.24	0.29	0.39
1	0.5	0.75	0.84	0.78	1.05	0.16	0.22	-0.05	0.2	0.3	0.5
1	0.9	0.25	0.99	0.98	1	0.01	0.01	0	0.03	0.04	0.04
1	0.9	0.50	0.98	0.98	1	0.02	0.02	0	0.03	0.04	0.05
1	0.9	0.75	0.98	0.97	1	0.02	0.03	0	0.03	0.04	0.06
0.5	0.5	0.25	0.49	0.46	0.49	0	0.04	0.01	0.32	0.36	0.39
0.5	0.5	0.50	0.54	0.45	0.48	-0.04	0.05	0.02	0.27	0.4	0.45
0.5	0.5	0.75	0.59	0.43	0.47	-0.09	0.07	0.03	0.23	0.51	0.59
0.9	0.9	0.25	0.9	0.9	0.9	0	0	0	0.04	0.04	0.04
0.9	0.9	0.50	0.92	0.89	0.9	-0.02	0.01	0	0.03	0.05	0.05
0.9	0.9	0.75	0.93	0.89	0.9	-0.03	0.01	0	0.03	0.06	0.06

We see from Table 10 that when reaching the full sample, we will have little bias in the CE when adjusting for the conditional bias from decision making. We also see that in terms of bias the CE is superior to both the MLE and CE-, where this conditioning was not adjusted for.

5.2.2 | Conclusive before reaching sample size

In this section we will look at simulation results where the final sample size is not reached. Hence the meta-analysis will be conclusive before reaching the full sample size. In this scenario the CE and PCE will not be the same. As the CE was found to be superior to the CE- in terms of bias, we will only compare PCE to CE and MLE. The simulation results are presented in Table 11 and are discussed in the next subsection.

TABLE 11 Simulation results based on 10000 meta-analyses where the true intervention effect is $\theta = \delta$ and the analysis becomes conclusive before reaching the full sample size. The first three columns are simulation settings. The remaining columns are the simulation results where bias is the average point estimate minus the true effect θ . IF_1 stands for the information fraction of the first study and translates to the fraction of participants in the first study compared to the full participant size. The event probability p_C is set to 0.2.

θ	δ	IF_1	Average point estimate			Bias			Standard deviation		
			MLE	CE	PCE	MLE	CE	PCE	MLE	CE	PCE
0.5	0.5	25	0.23	0.18	0.42	0.27	0.32	0.08	0.33	0.46	0.63
0.5	0.5	50	0.27	0.14	0.46	0.23	0.36	0.04	0.24	0.45	0.61
0.5	0.5	75	0.33	0.12	0.52	0.17	0.38	-0.02	0.16	0.39	0.52
0.9	0.9	25	0.85	0.83	0.89	0.05	0.07	0.01	0.03	0.04	0.07
0.9	0.9	50	0.86	0.8	0.9	0.04	0.1	0	0.02	0.04	0.08
0.9	0.9	75	0.88	0.77	0.9	0.02	0.13	0	0.01	0.05	0.08

6 | CONCLUSION

In Table 10, we find for meta-analyses reaching the full sample size which had an initial promising meta-analysis, the conditional estimator (CE) conditioning on $Z_1^{promising}$ is unbiased, whereas both the MLE and CE- where we do not condition on $Z_1^{promising}$ have bias. That CE- is more biased than the MLE when $\theta = 1$ is expected as the CE-, when at the full sample, will adjust the point estimate away from the null¹⁰. Using CE and CE- involves a moderate increase in the standard error compared to the MLE.

In Table 11, we investigated meta-analyses becoming conclusive before reaching the full sample size with an initial promising meta-analysis. Here the PCE can be compared with the CE and MLE. Both the CE and PCE are conditioning on $Z_1^{promising}$. Given the results from Table 11, we find that the PCE performs better than the CE and MLE in terms of bias. For $\theta = 1$ where $\delta \neq 1$, all three estimators would have problems with bias as an early conclusion in this scenario equates to conditioning on a type-I-error. This scenario is further very unlikely as described in Table 9. Furthermore, the sampling mechanism has the effect of making the z-scores from the first analyses are likely to be further away from the null, as the meta-analysis must stop at the second stage. This is not achievable with a first analysis too close to the null value. Regardless of the technical reasons for the bias, in general all estimators will have problems with conditioning on stopping for rejection of the null when the null is true. For this reason simulations involving $\theta = 1$ conditioning on early stopping have been excluded in previous studies¹² and we have not included the result of these simulations. In terms of variance we observe the same behavior as before with the MLE having the lowest variance and the CE having again a moderate increase. The PCE has the largest variance in the scenarios presented but the smallest bias.

7 | DISCUSSION

Conditional bias from decision making affects the naive point estimate pushing it towards a greater effect than the true intervention effect. We found that conditional estimators, inspired by single group sequential trials, can be used to remove this bias. We found that the PCE performs the best in terms of stability and bias. The instability of the CE was especially clear in the application. The performance of PCE, CE and MLE was assessed via simulation studies.

We recommend using the estimate from PCE as a sensitivity analysis, when the conductor of the meta-analysis is evaluating whether there is a risk of decision-making bias being present in the meta-analysis. Hence we recommend it to be evaluated together with the naive meta-analysis point estimate when one suspects conditional bias. This type of bias is important to consider when subsequent studies have been justified by earlier studies being promising and are then added to the existing evidence. Using the method as a sensitivity analysis, the adjusted estimate provides an unbiased estimate of the point estimate under complete adherence to the decision-making boundaries defined in the estimation. In this paper, we have only looked at the scenario where one continues the analysis if the observed z-score is between 0 and -1.96. However, the decision criteria can easily be changed in our framework to consider other limits. Using e.g. the minimal clinical value as one of the limits might be of more relevance in some scenarios. The method presented is robust towards other decision criteria. However, continuing significant analyses is not an option in our framework as we assume that the updated meta-analysis does not continue after a significant result. In this scenario it might be of more use to change the purpose of the meta-analysis to a Living Systematic Review which is less focused on the decision-making context²⁹.

The LIFT study is another practical example of potential conditional bias from decision making. This study was justified by a promising but not significant meta-analysis³⁰. Here the effect of lactoferrin, an antimicrobial protein, as a supplement compared to a diet without lactoferrin was investigated for lowering the risk of multiple outcomes but especially late-onset sepsis for preterm and low birth weight infants. As the LIFT study itself did not have sufficient power to answer all three hypotheses of interest, the study was added to the pre-existing meta-analysis, hence making the meta-analysis sequential. The LIFT study provides an additional example of a cumulative meta-analysis in which the methods developed here would be appropriate for adjusting for the conditional bias arising from decision making.

Another paper has mentioned the usefulness of creating an adjustment estimator. In⁶ it is mentioned that development of an appropriate bias adjustment should be possible under a explicit decision strategy. This is what we have created in this paper. However, it also mentioned that development of such a strategy requires the combined efforts of statisticians and decision-makers. For an explicit strategy, the result of the PCE might be evaluated equally important as the naive point estimate. For scenarios where the strategy is less explicit, the PCE might contribute more as a sensitivity analysis.

How to cite this article: Soerensen A. L., and Marschner I. C.. Adjusting for conditional bias in an updated meta-analysis arising from a decision to update. *Statistics in Medicine*. 2023;00(00):1–18.

APPENDIX

A ADJUSTED SEQUENTIAL META-ANALYSIS USING A SEQUENTIAL TESTING SCHEME

In this section we will continue investigating the behavior of the MLE, CE and PCE using the cord clamping example. Compared to the analysis presented in Section 4.2, we will see how the estimators behave when the MLE is closer to the boundary.

Had the meta-analyses been prospective, a sequential design would have controlled the type-I-error risk when continuously updating the meta-analysis. This is not the case for the cord clamping case study. However, we can investigate whether applying a sequential design would have changed whether we would have stopped the meta-analysis after adding the APTS trial and whether a different set of boundaries changes the CE and PCE.

Trial sequential analysis (TSA) is one way to create a sequential testing set up for meta-analysis⁷. We can calculate sequential stopping boundaries using the previously calculated RIS as the final sample size. We choose to use the Lan and DeMets version of O’Brien-Fleming α spending boundaries¹⁴. We modify the boundaries as we assume that APTS would not be planned unless the first meta-analysis was promising. Thus, we adjust the stopping boundaries to be 0 for the first boundary value in the direction of greater risk of in-hospital mortality when using delayed cord clamping. Table A1 shows the boundaries that we will consider.

TABLE A1 Lan and DeMets O’Brien-Fleming stopping boundaries. First boundary corresponds to the first meta-analysis, the second stopping boundary is for the update with the APTS study and the last is then hypothetically reaching the required information size.

	First boundary	Second boundary	Potential third boundary
Upper boundary	0	2.1975	2.0339
Lower boundary	-5.0349	-2.1975	-2.0339

We see that the meta-analysis can be stopped after updating with APTS with an observed z -value at -2.26. This is visualised in Figure 3 right plot, where the stopping boundary is just crossed at the second meta-analysis. As the meta-analysis stopped we compute the MLE, CE and PCE. The point estimates are found in Table A2.

TABLE A2 Point estimates.

	log(RR)	RR	log(CE)	CE	log(PCE)	PCE	λ^*
Point estimates	-0.3403	0.7116	2.1304	8.4182	-0.0385	0.9623	0.8588

We see that the CE estimates the RR to 8.42 which way over in the other direction than the MLE and the direction for which the meta-analysis stopped. The PCE satisfies that the adjusted point estimate is in the direction of which the trial stopped. When the stopping is extremely close to the boundary as in this example the CE is getting pushed towards ∞ and PCE is pushed towards the null value.

B CHANGING THE INFORMATION OF THE FINAL ANALYSIS

In the cord clamping case study, the first meta-analysis only accounted for approximately 20% of the information required to have a well-powered study of 80%. As an illustrative exercise in this section we consider hypothetical analyses with a greater amount information at the first meta-analysis. The reason being, that as we are conditioning on the stopping time T , we might imply an extreme loss of information. In the cord clamping application, the first meta-analysis accounts for a small fraction of the information compared to the updated meta-analysis with APTS. Hence the information and efficiency loss might not be as big for this particular application. We wish to see what happens to the estimates calculated used CE and PCE when the first

meta-analysis accounts for a specific fraction of the combined information. We can investigate this by adding more information to the first meta-analysis. Splitting ATPS in two parts, we can add one part to the first meta-analysis and let the other part be the added trial used for updating the meta-analysis. By varying the size of two parts, the implication of the size of the first meta-analysis can be investigated. Scenarios where respectively 50% and 75% of required information is placed in the first meta-analysis, will be investigated.

APTS had 782 participants in the control group and 784 participants in the intervention group. We are going to assume that there was no dependence between the timing of enrollment and the number of events. We split the APTS into two studies, adding the first study to the initial first meta-analysis to see how the information size of the first meta-analysis affects the estimates.

Table B3 shows the estimates, when respectively 50% and 75% of the required information size was allocated to the first meta-analysis. We find that the CE and PCE moves towards the naive estimate as the information fraction increases of the previous meta-analysis.

TABLE B3 Estimates (RR) from the MLE, CE and PCE when 50% and 75% of the information is spent on the first meta-analysis.

	MLE	CE	PCE	λ^*
50% information at first meta-analysis	0.7116	2.2445	0.902	0.8675
75% information at first meta-analysis	0.7116	1.524	0.8216	0.9178

C ADDITIONAL RESULTS

Additional simulation results in the scenario of reaching the full sample size.

TABLE C4 Simulation results based on 10000 meta-analyses where the true intervention effect $\theta = 1.1$. The first three columns are simulation settings. The remaining columns are the simulation results where bias is the average point estimate minus the true effect θ . IF_1 stands for the information fraction of the first study and translates to the fraction of participants in the first study compared to the full participant size. The event of probability p_C is set to 0.2.

θ	δ	IF_1	Average point estimate			Bias			Standard deviation		
			MLE	CE-	CE	MLE	CE-	CE	MLE	CE-	CE
1.1	0.5	0.25	0.99	0.98	1.11	0.11	0.12	-0.01	0.26	0.28	0.33
1.1	0.5	0.50	0.92	0.91	1.12	0.17	0.19	-0.02	0.23	0.27	0.38
1.1	0.5	0.75	0.88	0.84	1.16	0.22	0.26	-0.06	0.19	0.27	0.48
1.1	0.9	0.25	1.06	1.07	1.1	0.04	0.03	0	0.03	0.04	0.04
1.1	0.9	0.50	1.04	1.05	1.1	0.06	0.05	0	0.03	0.04	0.05
1.1	0.9	0.75	1.01	1.02	1.11	0.09	0.08	-0.01	0.02	0.03	0.07

REFERENCES

1. Kleinert S, Benham L, Collingridge D, Summerskill W, Horton R. Further emphasis on research in context. *The Lancet*. 2014;384:2176-2177.
2. CONSORT group *CONSORT 2010*. (accessed Sep. 13 2021).
3. Chalmers I, Bracken MB, Djulbegovic B, et al. How to increase value and reduce waste when research priorities are set. *The Lancet*. 2014;383(9912):156-165. doi: [https://doi.org/10.1016/S0140-6736\(13\)62229-1](https://doi.org/10.1016/S0140-6736(13)62229-1)
4. Ellis SP, Stewart JW. Temporal Dependence and Bias in Meta-Analysis. *Communications in Statistics - Theory and Methods*. 2009;38(15):2453-2462. doi: 10.1080/03610920802562772
5. terSchure J, Grünwald P. Accumulation Bias in meta-analysis: the need to consider time in error control. *F1000Research*. 2019;8:962. doi: 10.12688/f1000research.19375.1
6. Kulinskaya E, Huggins R, Dogo SH. Sequential Biases in Accumulating Evidence. *Research Synthesis Methods*. 2015;7(3):294-305. doi: 10.1002/jrsm.1185
7. 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(1):64-75. doi: 10.1016/j.jclinepi.2007.03.013

8. Higgins JPT, Whitehead A, Simmonds M. Sequential Methods for Random-Effects Meta-Analysis. *Statistics in Medicine*. 2010;30(9):903-921. doi: 10.1002/sim.4088
9. Jennison C, Turnbull B. *Group sequential tests with applications to clinical trials*. Chapman and Hall/CRC Interdisciplinary StatisticsUK United Kingdom: Chapman and Hall, 1999.
10. Fan XF, DeMets DL, Lan KKG. Conditional Bias of Point Estimates Following a Group Sequential Test. *Journal of Biopharmaceutical Statistics*. 2004;14(2):505-530. doi: 10.1081/bip-120037195
11. Liu A, Troendle JF, Yu KF, Yuan VW. Conditional Maximum Likelihood Estimation Following a Group Sequential Test. *Biometrical Journal*. 2004;46(6):760-768. doi: 10.1002/bimj.200410076
12. Marschner IC, Schou M, Martin AJ. Estimation of the Treatment Effect Following a Clinical Trial That Stopped Early for Benefit. *Statistical Methods in Medical Research*. 2022;31(12):2456-2469. doi: 10.1177/09622802221122445
13. Denne JS. Estimation Following Extension of a Study on the Basis of Conditional Power. *Journal of Biopharmaceutical Statistics*. 2000;10(2):131-144. doi: 10.1081/bip-100101018
14. Lan KKG, DeMets DL. Discrete Sequential Boundaries for Clinical Trials. *Biometrika*. 1983;70(3):659. doi: 10.2307/2336502
15. Whitehead A. *Meta-analysis of Controlled Clinical Trials*. 2002 John Wiley & Sons, Ltd., 2002.
16. Turner RM, Jackson D, Wei Y, Thompson SG, Higgins JPT. Predictive Distributions for Between-Study Heterogeneity and Simple Methods for Their Application in Bayesian Meta-Analysis. *Statistics in Medicine*. 2014;34(6):984-998. doi: 10.1002/sim.6381
17. Spence GT, Steinsaltz D, Fanshawe TR. A Bayesian Approach To Sequential Meta-analysis. *Statistics in Medicine*. 2016;35(29):5356-5375. doi: 10.1002/sim.7052
18. Thorlund K, Engstrøm J, Wetterslev J, Brok J, Imberger G, Gluud C. *User manual for trial sequential analysis (TSA)*. Copenhagen Trial Unit, Centre for Clinical Intervention Research; Copenhagen, Denmark: 2011.
19. Pogue JM, Yusuf S. Cumulating Evidence From Randomized Trials: Utilizing Sequential Monitoring Boundaries for Cumulative Meta-Analysis. *Controlled Clinical Trials*. 1997;18(6):580-593. doi: 10.1016/s0197-2456(97)00051-2
20. Kulinskaya E, Wood J. Trial Sequential Methods for Meta-Analysis. *Research Synthesis Methods*. 2013;5(3):212-220. doi: 10.1002/jrsm.1104
21. Strickland PAO, Casella G. Conditional Inference Following Group Sequential Testing. *Biometrical Journal*. 2003;45(5):515-526. doi: 10.1002/bimj.200390029
22. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; Vienna, Austria: 2023.
23. Fogarty M, Osborn DA, Askie L, et al. Delayed Vs Early Umbilical Cord Clamping for Preterm Infants: a Systematic Review and Meta-Analysis. *American Journal of Obstetrics and Gynecology*. 2018;218(1):1-18. doi: 10.1016/j.ajog.2017.10.231
24. Rabe H, Díaz-Rossello JL, Duley L, Dowswell T. Effect of timing of umbilical cord clamping and other strategies to influence placental transfusion at preterm birth on maternal and infant outcomes. *Cochrane Database of Systematic Reviews*. 2012(8). doi: 10.1002/14651858.CD003248.pub3
25. Tarnow-Mordi W, Morris J, Kirby A, Robledo K, Askie L, al eRB. Delayed Versus Immediate Cord Clamping in Preterm Infants. *New England Journal of Medicine*. 2017;377(25):2445-2455. doi: 10.1056/nejmoa1711281
26. Rabe H, Gyte GM, Díaz-Rossello JL, Duley L. Effect of Timing of Umbilical Cord Clamping and Other Strategies To Influence Placental Transfusion At Preterm Birth on Maternal and Infant Outcomes. *Cochrane Database of Systematic Reviews*. 2019;2019(9):nil. doi: 10.1002/14651858.cd003248.pub4
27. Tarnow-Mordi WO, Duley L, Field D, et al. Timing of cord clamping in very preterm infants: more evidence is needed. *American Journal of Obstetrics and Gynecology*. 2014;211(2):118-123. doi: https://doi.org/10.1016/j.ajog.2014.03.055
28. Jackson D, Turner R. Power Analysis for Random-Effects Meta-Analysis. *Research Synthesis Methods*. 2017;8(3):290-302. doi: 10.1002/jrsm.1240
29. Simmonds M, Salanti G, McKenzie J. Living Systematic Reviews: 3. Statistical Methods for Updating Meta-Analyses. *Journal of Clinical Epidemiology*. 2017;91(nil):38-46. doi: 10.1016/j.jclinepi.2017.08.008
30. Tarnow-Mordi WO, Abdel-Latif ME, Martin A, Pammi M, Robledo K, al ePM. The Effect of Lactoferrin Supplementation on Death Or Major Morbidity in Very Low Birthweight Infants (LIFT): a Multicentre, Double-Blind, Randomised Controlled Trial. *The Lancet Child & Adolescent Health*. 2020;4(6):444-454. doi: 10.1016/s2352-4642(20)30093-6

A. RTSA manual

This appendix contains the documentation for the R package **RTSA**. The package is available online from the Comprehensive R Archive Network (CRAN):

A. L. Soerensen et al. (2023b). *RTSA: 'Trial Sequential Analysis' for Error Control and Inference in Sequential Meta-Analyses*. R package version 0.2.1. URL: <https://github.com/AnneLyng/RTSA>

The package is a translation and an update of the original Trial Sequential Analysis software implemented in java. All computational functions are translated and re-written to R by the candidate. The graphical elements of the package is written by Markus Harboe Olsen and the candidate.

Package ‘RTSA’

August 30, 2023

Type Package

Title 'Trial Sequential Analysis' for Error Control and Inference in Sequential Meta-Analyses

Version 0.2.1

Description Frequentist sequential meta-analysis based on 'Trial Sequential Analysis' (TSA) in programmed in Java by the Copenhagen Trial Unit (CTU). The primary function is the calculation of group sequential designs for meta-analysis to be used for planning and analysis of both prospective and retrospective sequential meta-analyses to preserve type-I-error control under sequential testing. 'RTSA' includes tools for sample size and trial size calculation for meta-analysis and core meta-analyses methods such as fixed-effect and random-effects models and forest plots. TSA is described in Wetterslev et. al (2008) <[doi:10.1016/j.jclinepi.2007.03.013](https://doi.org/10.1016/j.jclinepi.2007.03.013)>. The methods for deriving the group sequential designs are based on Jennison and Turnbull (1999, ISBN:9780849303166).

License GPL (>= 2)

URL <https://github.com/AnneLyng/RTSA>

BugReports <https://github.com/AnneLyng/RTSA/issues>

Imports stats, metafor, ggplot2, scales, Rcpp (>= 0.11.0)

LinkingTo Rcpp

Encoding UTF-8

Depends R (>= 3.5.0)

LazyData true

RoxygenNote 7.2.3

Suggests gsDesign, CompQuadForm, dplyr, kableExtra, rmarkdown, knitr, bookdown, gridExtra

VignetteBuilder knitr, bookdown

NeedsCompilation yes

Author Anne Lyngholm Soerensen [aut, cre, trl],
 Markus Harboe Olsen [aut, ctr],
 Theis Lange [ctr],
 Christian Gluud [ctr]

Maintainer Anne Lyngholm Soerensen <lynganne@gmail.com>

Repository CRAN

Date/Publication 2023-08-30 09:10:21 UTC

R topics documented:

boundaries	2
coronary	4
eds	5
inference	6
metaanalysis	7
minTrial	9
perioOxy	11
plot.boundaries	12
plot.metaanalysis	12
plot.RTSA	13
ris	14
RTSA	16
Index	21

boundaries	<i>Boundaries for group sequential designs</i>
------------	--

Description

Calculates alpha- and potentially beta-spending boundaries for group sequential designs for meta-analysis. Should be used for exploring how the different arguments affect the sequential design. The function is not intended to be used individually for Trial Sequential Analysis. For this purpose, we recommend RTSA().

Usage

```
boundaries(  
  timing,  
  alpha = 0.05,  
  beta = 0.1,  
  side = 2,  
  futility = "none",  
  es_alpha = "esOF",  
  es_beta = NULL,  
  type = "design",
```

```

    design_R = NULL,
    tol = 1e-09
)

```

Arguments

timing	Expected timings of interim analyses and final analysis as a vector consisting of values from 0 to 1.
alpha	The level of type I error as a percentage, the default is 0.05 corresponding to 5%.
beta	The level of type II error as a percentage, the default is 0.1 corresponding to 10%.
side	Whether a 1- or 2-sided hypothesis test is used. Defaults to 2. Options are 1 or 2.
futility	Futility boundaries added to design. Options are: none, non-binding and binding. Default is "none".
es_alpha	The error spending function for alpha-spending. Options are: "esOF" (Lan & DeMets version of O'Brien-Fleming boundaries), "esPoc" (Lan & DeMets version of Pocock boundaries), "HSDC" (Hwang Sihi and DeCani) and "rho" (rho family). Defaults to "esOF".
es_beta	The error spending function for beta-spending. For options see es_alpha. Defaults to NULL.
type	Whether the boundaries are used for design or analysis. We recommend only to use the boundaries() function with type equal to design. Defaults to design.
design_R	If type is analysis, a scalar for achieving the right amount of power is required. It is recommended not to use the boundaires() function with the setting type equal to analysis. Defaults to NULL.
tol	Tolerance level for numerical integration. Defaults to 1e-09.

Value

A boundaries object which includes:

inf_frac	Timing of interim analyses and final analysis. Potentially modified if type = "analysis".
org_inf_frac	Original timing. If type = "design".
alpha_ubound	Upper alpha-spending boundaries
alpha_lbound	Lower alpha-spending boundaries
alpha	As input
alpha_spend	List of cumulative and incremental spending
delta	Drift parameter
design_R	If type = "analysis" it is the scalar for correct power in the design. Else NULL.

info	List of the information as the squareroot of the information increments and the squareroot of the cumulative information
beta_ubound	Upper beta-spending boundaries
beta_lbound	Lower beta-spending boundaries
root	Scalar for achieving correct power
beta_spend	List of cumulative and incremental spending
pwr	List of probabilities for rejecting the null under the sample size settings being true at each analysis and the sum.
tIe	List of probabilities for type-I-error at each analysis and the sum
side	As input
beta	As input
es_alpha	As input
es_beta	As input
type	As input
futility	As input

Examples

```
boundaries(timing = c(0.25, 0.5, 0.75, 1), alpha = 0.05, beta = 0.1,
  side = 2, futility = "non-binding", es_alpha = "es0F", es_beta = "es0F")
```

coronary

Dataset of trials investigating the intensity of statin therapy on the risk of myocardial infarction or coronary death

Description

A dataset containing trials investigating myocardial infarction or coronary death among patients with acute coronary syndromes or chronic coronary artery disease of statin therapy intensity. The trials compared low intensities of statin to higher intensities.

Usage

```
coronary
```

Format

A data frame with 4 rows and 5 variables:

- study** Name of first author of the trial
- eI** Number of events in the intervention group
- nI** Number of participants in the intervention group
- eC** Number of events in the control group
- nC** Number of participants in the control group

eds

Dataset of trials investigating the effect of carer on early supported discharge services

Description

A dataset containing trials investigating on the length of hospital stay when receiving early supported discharge (ESD) service versus conventional care. The outcome is length of initial hospital stay counted in days.

Usage

eds

Format

A data frame with 9 studies and 8 variables:

Details

- study. Name of the city of the study
- year. Year of the trial
- mI. Mean duration at hospital in intervention (ESD) group
- mC. Mean duration at hospital in control group
- sdI. Standard deviation of intervention (ESD) estimate
- sdC. Standard deviation of control estimate
- nI. Number of participants in the intervention (ESD) group
- nC. Number of participants in the control group

References

Fearon P, Langhorne P. Services for reducing duration of hospital care for acute stroke patients. Cochrane Database of Systematic Reviews 2012, Issue 9. Art. No.: CD000443. DOI: 10.1002/14651858.CD000443.pub3. Accessed 17 October 2022.

inference

*Inference calculations for sequential meta-analysis***Description**

Calculates point-estimates, p-values and confidence intervals. Computes naive inference and TSA-adjusted confidence intervals. If the meta-analysis crosses a alpha-spending boundary, a binding beta-spending boundary or reached the sequential RIS, stage-wise ordered inference is also calculated. This function is not supposed to be used individually for Trial Sequential Analysis (TSA). RTSA() is recommended for TSA.

Usage

```
inference(
  bounds,
  timing,
  ana_times,
  ma,
  fixed,
  org_timing,
  inf_type = "sw",
  conf_level = 0.95,
  final_analysis = FALSE,
  tol = 1e-15
)
```

Arguments

<code>bounds</code>	The boundaries for the analysis as calculated by the <code>boundaries()</code> function in RTSA.
<code>timing</code>	The timing of the studies relative to the sequential RIS. A vector consisting of values equal to the proportion of study participants out of the sequential RIS.
<code>ana_times</code>	The analysis times presented as a vector. Describes at which studies the meta-analyses were performed. If one expects that the meta-analysis was updated per study a vector from 1 to the number of studies included can be used.
<code>ma</code>	A metaanalysis object from the metaanalysis function.
<code>fixed</code>	Whether the analysis is for fixed-effect or random-effects meta-analysis. Options are TRUE (meta-analysis is fixed-effect) or FALSE (meta-analysis is random-effects).
<code>org_timing</code>	The timing of all included studies as a proportion of RIS and not sequential RIS.
<code>inf_type</code>	For now only option is "sw" (stage-wise). Type of inference used for point estimates, confidence intervals and p-values.
<code>conf_level</code>	The confidence interval level. Defaults to 0.95 which is 95%.
<code>final_analysis</code>	Whether or not the this analysis is considered the final analysis.
<code>tol</code>	The tolerance level. Set to 1e+09.

Value

A data.frame of cumulative meta-analysis results including stopping boundaries and a list of conditional sequential inference to be parsed to RTSA

`results_df` A data.frame containing information about: Cumulative test values, cumulative outcomes, timing of trials, stopping boundaries (`alpha_upper`, `alpha_lower`, `beta_upper`, `beta_lower`), naive confidence intervals, TSA-adjusted confidence intervals, cumulative p-values and standard deviations.

`seq_inf` If the meta-analysis crosses an alpha-spending boundary, a binding beta-spending boundary or reaches the required information size inference conditional on stopping is provided. A median unbiased estimate, lower and upper confidence interval, and p-value is calculated based on stage-wise ordering.

Examples

```
ma <- metaanalysis(data = perioOxy, outcome = "RR", mc = 0.8)
sts <- ma$ris$NR_D2$NR_D2_full
timing <- cumsum(perioOxy$I + perioOxy$C)/sts
bound_oxy <- boundaries(timing = timing, alpha = 0.05, beta = 0.2, side = 2,
                       futility = "none", es_alpha = "esOF")
inference(timing = bound_oxy$inf_frac, bounds = bound_oxy, ma = ma, fixed = FALSE,
ana_times = 1:length(timing), org_timing = timing)
```

 metaanalysis

Fixed-effect or random-effects meta-analysis

Description

Computes a fixed-effect or random-effects meta-analysis including heterogeneity statistics. If `mc` is specified, a retrospective sample and trial size is calculated.

Usage

```
metaanalysis(
  outcome,
  data,
  side = 2,
  alpha = 0.05,
  beta = 0.1,
  weights = "IV",
  re_method = "DL_HKSJ",
  tau_ci_method = "BJ",
  cont_vartype = "equal",
  mc = NULL,
  RRR = NULL,
  sd_mc = NULL,
```

```

    study = NULL,
    conf_level = 0.95,
    zero_adj = 0.5,
    ...
)

```

Arguments

outcome	Outcome metric for the studies. Choose between: MD (mean difference), RR (relative risk), RD (risk difference), or OR (odds ratio).
data	A data.frame containing the study results. The data set must contain a specific set of columns. These are respectively 'eI' (events in intervention group), 'eC' (events in control group), 'nC' (participants intervention group) or 'nI' (participants control group) for discrete data, or, 'mI' (mean intervention group), 'mC' (mean control group), 'sdI' (standard error intervention group), 'sdC' (standard error control group), 'nC' (participants intervention group) and 'nI' (participants control group) for continuous outcomes. Preferable also a 'study' column as an indicator of study.
side	Whether a 1- or 2-sided hypothesis test is used. Options are 1 or 2. Default is 2.
alpha	The level of type I error as a percentage, the default is 0.05 corresponding to 5%.
beta	The level of type II error as a percentage, the default is 0.1 corresponding to 10%. Not used unless a sample and trial size calculation is wanted.
weights	Method for calculating weights. Options are "MH" (Mantel-Haenzel and only optional for binary data) or "IV" (Inverse variance weighting). Default is "IV".
re_method	Methods are "DL" for DerSimonian-Laird or "DL_HKSJ" for DerSimonian-Laird with Hartung-Knapp-Sidik-Jonkman adjustment. Default is "DL_HKSJ".
tau_ci_method	Methods for computation of confidence interval for heterogeneity estimate tau. Calls rma.uni from the metafor package. Options are "BJ" and "QP". Default is "BJ"
cont_vartype	Variance type for continuous outcomes. Choices are "equal" (homogeneity of treatment group variances) or "non-equal" (heterogeneity of treatment group variances). Default is "equal".
mc	Minimum clinically relevant value. Used for sample and trial size calculation.
RRR	Relative risk reduction. Used for binary outcomes with outcome metric RR. Argument mc can be used instead. Must be a value between 0 and 1.
sd_mc	The expected standard deviation. Used for sample and trial size calculation for mean differences.
study	Optional vector of study IDs. If no study indicator is provided in 'data', a vector of study indicators e.g. names.
conf_level	Confidence interval coverage
zero_adj	Zero adjustment for null events in binary data. Options for now is 0.5. Default is 0.5.
...	Additional variables.

Value

A metaanalysis object which is a list with 6 or 7 elements.

study_results	A data.frame containing study results which is information about the individual studies
meta_results	A data.frame containing the results of the meta-analysis such as the pooled estimate, its standard error, confidence interval and p-value
hete_results	A list containing statistics about heterogeneity.
metaPrepare	A list containing the elements used for calculating the study results.
synthesize	A list containing the elements used for calculating the meta-analysis results.
settings	A list containing the arguments used in the metaanalysis call.
ris	(Only when mc has been specified or meta-analysis is created as part of RTSA). List of sample size and trial size calculation. See documentation for ris.

Examples

```
### Basic uses
# Use perioOxy data from package and run meta-analysis with default settings
data(perioOxy)
metaanalysis(outcome = "RR", data = perioOxy, study = perioOxy$trial)

# Run same meta-analysis but with odds ratio as outcome metric, Mantel-Haenzel
# weights and DerSimonian-Laird for the variance estimate
metaanalysis(outcome = "OR", data = perioOxy, study = perioOxy$trial,
  weights = "MH", re_method = "DL")

# Run meta-analysis with mean difference as outcome metric
data(eds)
metaanalysis(outcome = "MD", data = eds)

### Retrospective sample size calculation
# minimal clinically relevant difference set to an odds ratio of 0.7.
ma <- metaanalysis(outcome = "OR", data = perioOxy, mc = 0.7)
ma$ris
```

minTrial

Minimum number of trials needed for a specific level of power

Description

Calculates minimum number of trials needed to achieve power in a meta-analysis with heterogeneity.

Usage

```

minTrial(
  outcome,
  mc,
  tau2,
  alpha,
  beta,
  side,
  pC = NULL,
  p1 = NULL,
  var_mc = NULL,
  var_random = NULL,
  trials = NULL
)

```

Arguments

outcome	Metric of interest, options include "RR" (relative risk), "OR" (odds ratio), "RD" (risk difference) and "MD" (mean difference).
mc	Minimal clinical relevant value provided as a numeric value. Such as 0.8 for e.g. an odds ratio of 0.8.
tau2	Heterogeneity estimate. Can be extracted from the metaanalysis() function.
alpha	The level of type I error as a percentage, the default is 0.05 corresponding to 5%.
beta	The level of type II error as a percentage, the default is 0.1 corresponding to 10%.
side	Whether a 1- or 2-sided hypothesis test is used. Options are 1 or 2.
pC	Probability of event in control group. Only used for outcomes "RR", "OR" and "RD".
p1	Probability of event in treatment group. Only used for outcome "RD".
var_mc	Variance of the estimated effect when outcome is "MD". Not required for outcome types "OR", "RR" or "RD".
var_random	Estimated variance from the random-effects meta-analysis. Used then a meta-analysis have already been made previously.
trials	Optional argument. Number of trials of interest for to provide the number of participants needed for that exact number of trials.

Value

Either a number (minimum required trials) or the minimum required required trials together with a matrix of required participants per trial given different number of trials.

Examples

```
# Minimum number of trials for a prospective meta-analysis
minTrial(outcome = "RR", pC = 0.5, mc = 0.7, tau2 = 0.05, alpha = 0.05,
beta = 0.1, side = 2)

# Minimum number of trials still needed for a retrospective meta-analysis
# Note that retrospective sample size calculations are prone to bias
ma <- metaanalysis(outcome = "RR", data = perioOxy)
ris(outcome = "RR", mc = 0.80, ma = ma, type = "retrospective", fixed = FALSE,
beta = 0.1, alpha = 0.05, side = 2)
```

perioOxy	<i>Dataset of RCTs investigating the effect of 80% perioperative oxygen vs. 30-35% perioperative oxygen on surgical site infection.</i>
----------	---

Description

A dataset containing data on seven trials which includes their number of events per treatment group, where intervention is 80% oxygen and control is 30-35% oxygen, number of participants in each treatment group and the year of the trial.

Usage

```
perioOxy
```

Format

A data frame with 7 rows and 6 variables:

study Name of first author of the trial

eI Number of events in the intervention group (80% oxygen)

nI Number of participants in the intervention group (80% oxygen)

eC Number of events in the control group (30-35% oxygen)

nC Number of participants in the control group (30-35% oxygen)

plot.boundaries *Plot of boundaries for group sequential designs*

Description

Plot of boundaries for group sequential designs

Usage

```
## S3 method for class 'boundaries'
plot(x, theme = "classic", ...)
```

Arguments

x	boundaries object
theme	Whether the theme is "classic" or "aussie"
...	Other arguments to plot.boundaries

Value

Plot. Either a plot for two- or one-sided testing.

Examples

```
bounds <- boundaries(timing = c(0.5,0.75, 1), alpha = 0.025, beta = 0.2,
side = 1, futility = "none", es_alpha = "esOF")
plot(x = bounds)
```

plot.metaanalysis *Forestplot for metaanalysis object.*

Description

Forestplot for metaanalysis object.

Usage

```
## S3 method for class 'metaanalysis'
plot(x, type = "both", xlims = NULL, ...)
```

Arguments

x	metaanalysis object from the RTSA package.
type	Define whether or not both fixed-effect and random-effects meta-analysis results should be printed on the plot. Options are: "fixed", "random" or "both". Default is "both".
xlims	Set default limits on the outcome scale. Default is NULL.
...	Additional arguments

Examples

```
# Example with OR
ma <- metaanalysis(data = coronary, outcome = "OR")
plot(ma)

# Example with RR
ma <- metaanalysis(data = perioOxy, outcome = "RR")
plot(ma)

# Example with MD
ma <- metaanalysis(data = eds, outcome = "MD")
plot(ma, type = "random")
```

plot.RTSA

Plot RTSA object. Returns the R version of the original TSA plot.

Description

Plot RTSA object. Returns the R version of the original TSA plot.

Usage

```
## S3 method for class 'RTSA'
plot(x, model = "random", type = "classic", theme = "classic", ...)
```

Arguments

x	RTSA object
model	Whether a fixed- or random-effects meta-analysis should be used. Defaults to random.
type	Should Z-scores (classic) or outcome values (outcome) be plotted.
theme	Whether the theme is traditional TSA (classic) or modern (modern)
...	Other arguments to plot.RTSA

Value

Plot. Either a plot for two sided testing or one-sided

Examples

```
data(peri0xy)
outRTSA <- RTSA(type = "analysis", data = peri0xy, outcome = "RR", mc = 0.8,
  side = 2, alpha = 0.05, beta = 0.2, fixed = FALSE, es_alpha = "es0F", design = NULL)
plot(x = outRTSA)
```

ris

*Calculate required sample and trials size.***Description**

Calculate required sample and trials size.

Usage

```
ris(
  outcome,
  mc,
  side = 2,
  alpha = 0.05,
  beta = 0.1,
  fixed = TRUE,
  sd_mc = NULL,
  pC = NULL,
  p1 = NULL,
  ma = NULL,
  tau2 = NULL,
  I2 = NULL,
  D2 = NULL,
  type = "prospective",
  trials = NULL,
  RTSA = FALSE,
  ...
)
```

Arguments

outcome	Choose between: "MD" (mean difference), "RR" (relative risk), "OR" (odds ratio) or "RD" (risk difference).
mc	Minimum clinical relevant effect. For "OR" or "RR" set to natural scale, not log scale.
side	Test type. Set to 1 or 2 depending on the test being 1- or 2-sided.
alpha	The level of type I error as a percentage, the default is 0.05 corresponding to 5%.

beta	The level of type II error as a percentage, the default is 0.1 corresponding to 10%.
fixed	Should sample size be based on a fixed-effect (TRUE) or random-effects (FALSE) model. Defaults to TRUE.
sd_mc	Standard deviation of estimated effect. Only needed when outcome type is "MD".
pC	Probability of event in control group. Only needed when outcome type is "OR", "RR" or "RD".
p1	Probability of event in treatment group. Only needed when outcome type is "RD".
ma	An optional metaanalysis object. Required for retrospective sample size calculations.
tau2	The value of the heterogeneity. Use when estimating the sample size under a random effects model. If data is provided, the estimated heterogeneity is used instead.
I2	Optional argument. Inconsistency.
D2	Optional argument. Diversity.
type	Whether the type of calculation is for "prospective" meta-analysis or "retrospective" meta-analysis. If the type is retrospective, one should add a meta-analysis object to the function. See argument ma.
trials	Optional numeric argument. If one is interested in a specific number of trials.
RTSA	Whether the ris function was called via the RTSA function. Purely operational argument.
...	additional arguments

Value

A list of up to 6 elements:

settings	A list containing the arguments provided to the ris function.
NF	The total number of required participants in a fixed-effect meta-analysis if type is prospective. Contains a list if the type is retrospective, where NF is the additional required number of participants and NF_full is the total required number of participants.
NR_tau	A list containing: minTrial the minimum number of trials. nPax a matrix containing four possible number of trials with the number of participants per trial and total number of participants. tau2 the estimate used for the calculation. Might contain NR_tau_l1 and NR_tau_u1 which contain the same three elements. NR_tau_l1 is based on the lower value in the confidence interval of tau2. NR_tau_u1 is based on the upper value in the confidence interval for tau2. If the type is prospective the numbers are the total required. If the type is retrospective the numbers are the additional required.
NR_D2	The total number of required participants in a random-effects meta-analysis adjusted by diversity (D2) if type is prospective. Contains a list if the type is retrospective, where NR_D2 is the additional required number of participants and NR_D2_full is the total required number of participants.

NR_I2 The total number of required participants in a random-effects meta-analysis adjusted by inconsistency (I2) if type is prospective. Contains a list if the type is retrospective, where NR_I2 is the additional required number of participants and NR_I2_full is the total required number of participants.

Examples

```
# Sample and trial size calculation for prospective meta-analysis
ris(outcome = "RR", mc = 0.8, pC = 0.12, fixed = TRUE, alpha = 0.05,
beta = 0.1, side = 2)

# Additional sample and trial size calculation for retrospective meta-analysis
# It is calculated directly from the metaanalysis() function
data("perioOxy")
ma <- metaanalysis(outcome = "RR", data = perioOxy, mc = 0.8, beta = 0.2)
ma$ris
# Or by using the two functions in sequence
ma <- metaanalysis(outcome = "RR", data = perioOxy)
ris(outcome = "RR", mc = 0.8, ma = ma, type = "retrospective", fixed = FALSE,
beta = 0.2, alpha = 0.05, side = 2)
```

RTSA

R version of Trial Sequential Analysis. Used for designing and analysing sequential meta-analyses.

Description

R version of Trial Sequential Analysis. Used for designing and analysing sequential meta-analyses.

Usage

```
RTSA(
  type = "design",
  outcome = NULL,
  side = 2,
  alpha = 0.05,
  beta = 0.1,
  futility = "none",
  es_alpha = "esOF",
  es_beta = NULL,
  timing = NULL,
  data = NULL,
  design = NULL,
  ana_times = NULL,
  fixed = FALSE,
  mc = NULL,
  RRR = NULL,
  sd_mc = NULL,
```

```

    pC = NULL,
    weights = "IV",
    re_method = "DL_HKSJ",
    tau_ci_method = "BJ",
    gamma = NULL,
    rho = NULL,
    study = NULL,
    cont_vartype = "equal",
    zero_adj = 0.5,
    tau2 = NULL,
    I2 = NULL,
    D2 = NULL,
    trials = NULL,
    final_analysis = NULL,
    inf_type = "sw",
    conf_level = 0.95,
    random_adj = "tau2",
    power_adj = TRUE,
    ...
)

```

Arguments

type	Type of RTSA. Options are "design" or "analysis".
outcome	Outcome metric. Options are: RR (risk ratio/relative risk), OR (odds ratio), RD (risk difference) and MD (mean difference).
side	Whether a 1- or 2-sided hypothesis test is used. Options are 1 or 2. Default is 2.
alpha	The level of type I error as a percentage, the default is 0.05 corresponding to 5%.
beta	The level of type II error as a percentage, the default is 0.1 corresponding to 10%.
futility	Futility boundaries added to design. Options are: none, non-binding and binding. Default is "none".
es_alpha	The spending function for alpha-spending. Options are: esOF (Lan & DeMets version of O'Brien-Fleming), esPoc (Lan & DeMets version of Pocock), HSDC (Hwang Sihi and DeCani) and rho (rho family).
es_beta	The spending function for beta-spending. For options see es_alpha.
timing	Expected timings of interim analyses when type = "design". Defaults to NULL.
data	A data.frame containing the study results. The data set must containing a specific set of columns. These are respectively 'eI' (events in intervention group), 'eC' (events in control group), 'nI' (participants intervention group) or 'nI' (participants control group) for discrete data, or, 'mI' (mean intervention group), 'mC' (mean control group), 'sdI' (standard error intervention group), 'sdC' (standard error control group), 'nI' (participants intervention group) and 'nI' (participants control group) for continuous outcomes. Preferable also a 'study' column as an indicator of study.

design	RTSA object where type is design.
ana_times	An optional vector of analysis times. Used if the sequential analysis is not done for all studies included in the meta-analysis.
fixed	Should only a fixed-effect meta-analysis be computed. Default is FALSE.
mc	Minimal clinical relevant outcome value
RRR	Relative risk reduction. Used for binary outcomes with outcome metric RR. Argument mc can be used instead. Must be a value between 0 and 1.
sd_mc	The expected standard deviation. Used for sample size calculation for mean differences.
pC	The expected probability of event in the control group. Used for sample size calculation for binary outcomes.
weights	Weighting method options include IV (inverse-variance) and MH (Mantel-Haenszel). Defaults to IV.
re_method	Method for calculating the estimate of heterogeneity, τ^2 , and the random-effects meta-analysis variance. Options are "DL" for DerSimonian-Laird and "DL_HKSJ" for the Hartung-Knapp-Sidik-Jonkman adjustment of the DerSimonian-Laird estimator.
tau_ci_method	Method for calculating confidence intervals for the estimated heterogeneity τ^2 . Options are "QP" for Q-profiling and "BJ" for Biggelstaff
gamma	Parameter for the HSDC error spending function.
rho	Parameter for the rho family error spending function.
study	An optional vector of study names and perhaps year of study. Defaults to NULL.
cont_vartype	For mean difference outcomes, do we expect the variance in the different groups to be "equal" or "non-equal".
zero_adj	Zero adjustment. Options for now is 0.5.
tau2	Heterogeneity estimate. Used for sample and trial size calculation. Defaults to NULL.
I2	Inconsistency estimate. Used for sample and trial size calculation. Defaults to NULL.
D2	Diversity estimate. Used for sample and trial size calculation. Defaults to NULL.
trials	Number of anticipated extra trials. Used for heterogeneity adjustment by tau2.
final_analysis	Whether or not the current analysis is the final analysis.
inf_type	Stopping time confidence interval. Options for now is sw (stage-wise).
conf_level	Confidence level on stopping time confidence interval.
random_adj	The sample size adjustment based on presence of heterogeneity. Options are "D2" (Diversity), "I2" (Inconsistency) and "tau2" (the heterogeneity estimate). Default is "tau2".
power_adj	Whether the sample size should be adjusted by the sequential design. Defaults to TRUE.
...	other arguments

Value

A RTSA object, a list of five elements:

settings	A list containing all of the settings used in the RTSA call. See Arguments.
ris	List containing sample and trial size calculations for a non-sequential meta-analysis. See documentation for ris function.
bounds	List of stopping boundaries, timing of trials and more. See documentation for boundaries function.
results	List of 3 to 7 elements. AIS Achieved information size. RIS Fixed-effect required information size for a non-sequential meta-analysis. SMA_RIS RIS adjusted for sequential analysis. HARIS Heterogeneity adjusted required information size for a non-sequential meta-analysis. SMA_HARIS HARIS adjusted for sequential analysis. results_df a data.frame of inference, see documentation for inference function. seq_inf a list of conditional inference, see documentation for inference function. metaanalysis A metaanalysis object, see documentation for metaanalysis function. design_df a data.frame containing the stopping boundaries and timings from the design.
warnings	List of warnings

Examples

```
## Not run:
### Retrospective sequential meta-analysis:
# A RRR of 20% is expected which gives mc = 1 - RRR = 0.8.
# No futility boundaries
data(periOxy)
RTSA(type = "analysis", data = periOxy, outcome = "RR", mc = 0.8, side = 2,
      alpha = 0.05, beta = 0.2, es_alpha = "esOF")

# Set binding futility boundaries
# And use Lan and DeMets' version of Pocock stopping boundaries
RTSA(type = "analysis", data = periOxy, outcome = "RR", mc = 0.8, side = 2,
      alpha = 0.05, beta = 0.2, es_alpha = "esOF", futility = "binding",
      es_beta = "esPoc")

# Set non-binding futility boundaries
RTSA(type = "analysis", data = periOxy, outcome = "RR", mc = 0.8, side = 2,
      alpha = 0.05, beta = 0.2, es_alpha = "esOF", futility = "non-binding",
      es_beta = "esPoc")

### Design a prospective sequential meta-analysis
# For continuous data without expected heterogeneity
RTSA(type = "design", outcome = "MD", mc = 5, sd_mc = 10, side = 1,
      timing = c(0.33, 0.66, 1), fixed = TRUE,
      alpha = 0.025, beta = 0.1, es_alpha = "esOF", futility = "non-binding",
      es_beta = "esPoc")

# For binary outcome
RTSA(type = "design", outcome = "RR", mc = 0.75, side = 1,
      timing = c(0.33, 0.66, 1), pC = 0.1, D2 = 0.1,
```

```
alpha = 0.025, beta = 0.2, es_alpha = "esOF", futility = "non-binding",
es_beta = "esOF")

# extract sample size calculation
out_rtisa <- RTSA(type = "design", outcome = "RR", mc = 0.75, side = 1,
timing = c(0.33, 0.66, 1), pC = 0.1, D2 = 0.1,
alpha = 0.025, beta = 0.2, es_alpha = "esOF", futility = "non-binding",
es_beta = "esOF")
out_rtisa$ris

# plot the design
plot(out_rtisa)

# update the design with data as it accumulates (here toy-data)
fake_data <- data.frame(eI = c(10,10), eC = c(13, 11), nI = c(750, 750),
nC = c(750,750))
RTSA(type = "analysis", design = out_rtisa, data = fake_data)

# plot the analysis
an_rtisa <- RTSA(type = "analysis", design = out_rtisa, data = fake_data)
plot(an_rtisa)

## End(Not run)
```

Index

* datasets

- coronary, 4
- eds, 5
- perio0xy, 11

boundaries, 2

coronary, 4

eds, 5

inference, 6

metaanalysis, 7

minTrial, 9

perio0xy, 11

plot.boundaries, 12

plot.metaanalysis, 12

plot.RTSA, 13

print.metaanalysis (metaanalysis), 7

print.ris (ris), 14

print.RTSA (RTSA), 16

ris, 14

RTSA, 16

B. Additional results

One of the original objectives of this PhD thesis was to compare Trial Sequential Analysis (TSA) to other methods for sequential meta-analysis in terms of type-I- and type-II-error control. In this section we will look at just TSA and see how well it controls the errors under various settings. These settings include different simulated scenarios such as when study results are homogeneous or heterogeneous. The following will present the different scenarios considered, the simulation scheme and the results.

Some of these results were presented at the meta-analysis section of the International Society of Clinical Biostatistics 44th conference in Milan Italy the 28th of August 2023 by the candidate.

We start this section with the settings of the simulation study which is followed by the simulation scheme. Two examples of a simulated sequential meta-analyses using the simulation scheme are then presented. The end of this section contains the simulation results and a discussion of the results.

B.1 Settings

We are going to investigate the level of the type-I- and type-II-errors of TSA under different scenarios using relative risks (RR). When we want to investigate the type-I-error, we will consider a true RR of 1. When we want to investigate the type-II-error, we will consider a true RR of 0.8. Table B.1 contains the simulation parameters and their values. The table also contains different sizes of between-trial-variation τ^2 . This means that the control of the errors are also going to be investigated under different levels of heterogeneity.

The simulation scheme is described next. The scheme describes when we consider to encounter a type-I- or type-II-error and how the simulation differs between a prospective and retrospective sequential meta-analysis. To illustrate the steps of the scheme, two examples of sequential meta-analyses are provided after the presentation.

B.2 Simulation scheme

The steps of the simulation is divided into three categories: Pre-simulation settings, sequential meta-analysis simulation, and sequential meta-analysis results. The first category is the preparation of the simulation which includes an initial sample size calculation. The second category simulates studies and calculates and updates the sequential meta-analysis until it reaches a stopping criteria. The third and final category collects the results and re-runs the sim-

Parameter	Values
Type of analysis	Prospective, retrospective
Statistic	RR (relative risk)
p_C (prob. of event in control group)	0.1
θ (true effect size)	0.8, 1
δ (minimal relevant clinical value)*	0.8
τ^2 (between trial-variation)	0, 0.05, 0.02
Maximum number of trials	50
α -spending function	Lan & DeMets' version of O'Brien-Fleming
Meta-analysis model	Fixed, Random
Trial size ($\#t$)	5, 15
Design accounts for heterogeneity (prospective)	Yes, No

Table B.1: Parameters and their values considered for the simulation studies in this section. *Used for initial sample size calculation.

ulation until the desired number of simulations has been reached. We are now ready to introduce the scheme:

Pre-simulation settings

1. Set a selection of the simulation parameter values illustrated in Table B.1.
2. Calculate an initial total sample size, denoted N , based on the minimal clinical relevant value δ and expected probability of event in the control group p_C . Given the total sample size N , we will assume a trial size of $N/(\#t)$ where $\#t$ is the initial number of trials as defined in Table B.1.
 - If the type of analysis is a prospective meta-analysis, the initial sample size calculation might include a minimal number of trials given the anticipated size of the heterogeneity. Update $\#t$ trial number to the minimal number of trials required. See Chapter 2 for information about the calculation of the number of trials.
 - If the type of analysis is retrospective, stick with the chosen value of $\#t$ from Table B.1.

Given the total sample size N , set the size of each simulated trial size to $N/(\#t)$.

3. If the type of analysis is a prospective meta-analysis, create a sequential design for the analysis. Using the RTSA package, it can look like this:

```
design_rtsa <-
  RTSA(
    type = "design",
    outcome = "RR",
    side = 2,
```

```

alpha = 0.05,
beta = 0.1,
es_alpha = "es0F", # O'Brien-Fleming boundaries
pC = 0.1, # prob. of event in the control group
# timing of the trials
timing = seq(1 / trials, 1, length.out = trials),
mc = 0.8, # minimal clinical relevant value
fixed = T # fixed-effect meta-analysis model used
)

```

The sample size N is then updated to the sample size calculated from the design which accounts for the sequential testing regime.

Sequential meta-analysis simulation

4. Given the sample size calculation and the simulation parameters, simulate the number of events in the two treatment arms.
5. Start with analysing one trial and calculate the z -score. If the design is prospective, evaluate the trial against the stopping boundaries from the design. If the design is retrospective, evaluate the trial against naive stopping boundaries.
 - If the test is inconclusive (no boundary has been crossed - or the sample size N has not been met), and the maximum number of trials have not been reached, add another trial and recalculate the z -score.
 - If the test is conclusive, go to step 8.
6. For retrospective meta-analyses, re-calculate the required sample size and denote it $N_{simulation}$ - if there is any presence of heterogeneity the sample size calculation will take this into account and use the method proposed by Kulinskaya et al. (2013) to calculate the sample size. See Chapter 2. Set $N = N_{simulation}$. Then calculate a retrospective sequential meta-analysis based on the observed data. Evaluate the new z -score against the newly calculated boundaries. For prospective meta-analyses evaluate the new z -score using the design.
 - If the test is inconclusive, and the maximum number of trials have not been reached, add another trial and recalculate the z -score.
 - If the test is conclusive or the maximum number of trials have been reached, go to step 8.
7. Redo step 6, unless all simulated trials have been used. In this case, simulate a new set of trials as per step 4.

Sequential meta-analysis results

8. Store the conclusion of the trial. If the meta-analysis stopped by crossing a boundary, it counts as a type-I-error under $\theta = 1$. If it does not, it counts as a type-II-error under $\theta \neq 1$.

9. Run step 4 through 8 10000 times.

Step 9 concludes the simulation. Examples of simulations using the steps of the scheme are shown in the next section.

B.3 Examples of simulated sequential meta-analyses

We will show two examples of simulated sequential meta-analyses, one is a prospective sequential meta-analysis and the other is a retrospective sequential meta-analysis.

Prospective sequential meta-analysis

The settings of the prospective sequential meta-analysis is shown in Table B.2. We are going to go through one simulation from this set-up step by step.

Parameter	Values
Type of analysis	Prospective
Statistic	RR (relative risk)
p_C (prob. of event in control group)	0.1
θ (true effect size)	1
δ (minimal relevant clinical value)*	0.8
τ^2 (between trial-variation)	0
Maximum number of trials	5
α -spending function	Lan & DeMets' version of O'Brien-Fleming
Meta-analysis model	Fixed
Trial size ($\#t$)	5
Design with heterogeneity	No

Table B.2: Parameters and their values considered for a specific simulation study. *Used for initial sample size calculation

Pre-simulation settings

We set the simulation parameters illustrated in Table B.2. Then an initial total sample size N is calculated based on the minimal clinical relevant value $\delta = 0.8$ and expected probability of event in the control group $p_C = 0.1$. The initial sample size is 8606.

As the meta-analysis is prospective, we add a design to the meta-analysis. The design is created via the RTSA package and creates boundaries and more:

```
design_rtsa <-
RTSA(
  type = "design",
  outcome = "RR",
  side = 2,
  alpha = 0.05,
  beta = 0.1,
  es.alpha = "esOF",
```

```

    pC = 0.1,
    timing = c(0.2, 0.4, 0.6, 0.8, 1),
    mc = 0.8,
    fixed = T
  )
design_rtsa

## Design with Trial Sequential Analysis was computed with the
## following settings:
##
## Boundaries for a 2-sided design with a type-I-error of 0.05,
## and type-II-error of 0.1.
## Futility is set to: none. Alpha-spending function: esOF.
## Beta-spending function: .
##
## The required information size is not adjusted by heterogeneity.
## The required information size is further increased with 2
## percent due to the sequential design. The total required
## information size is 8805.
##
## Timing, and boundaries:
## sma_timing upper lower
##      0.205 4.877 -4.877
##      0.409 3.357 -3.357
##      0.614 2.680 -2.680
##      0.818 2.290 -2.290
##      1.023 2.031 -2.031
## sma_timing is the ratio of the required sample for a
## sequential meta-analysis to a non-sequential meta-analysis
## sample size.
...

```

The design provides an updated estimate of the sample size as it adjusts according to the sequential design. The new sample size is 8805 and will continue to be 8805 throughout the analysis as using a prospective design means that sample size calculation does not get updated each time a new trial is added to the analysis. Hence the design will not change if there is presence of heterogeneity or the probability of event in the control group was different than anticipated.

Sequential meta-analysis simulation

The number of events in the two treatment arms are simulated for the 5 trials. Analysing the first trial and calculating the z -score to approximately 0, we have an inconclusive trial given the boundaries of the design is 4.88 and -4.88 per the above design. This means that we will continue the analysis.

Continuing the analysis, we get the following z -scores, acquired information sizes (AIS), required information sizes (RIS) and the estimated value of the heterogeneity $\hat{\tau}^2$ in Table B.3. As the setting in the simulation (see Table B.2) the heterogeneity is estimated to 0. Given the update of the meta-analysis,

# Trials	z -score	Boundary	AIS	RIS	$\hat{\tau}^2$
2	-0.4	-3.36	3520	8805	0
3	-1.02	-2.68	5280	8805	0
4	-0.52	-2.29	7040	8805	0
5	-0.15	-2.03	8800	8805	0

Table B.3: Outcome of each meta-analysis from one simulation of a prospective sequential meta-analysis.

the analysis concludes with not being able to reject the null hypothesis which aligns with the true value of the RR being 1.

We now continue with an example of retrospective sequential meta-analysis using the simulation scheme.

Retrospective sequential meta-analysis

The settings of the simulation is shown in Table B.4. We are going to go through one simulation step by step.

Parameter	Values
Type of analysis	Retrospective
Statistic	RR (relative risk)
p_C (prob. of event in control group)	0.1
θ (true effect size)	0.8
δ (minimal relevant clinical value)*	0.8
τ^2 (between trial-variation)	0.02
Maximum number of trials	50
α -spending function	Lan & DeMets' version of O'Brien-Fleming
Meta-analysis model	Random
Trial size ($\#t$)	15
Design with heterogeneity	No

Table B.4: Parameters and their values considered for a specific simulation study. *Used for initial sample size calculation

Pre-simulation settings

We set the simulation parameters illustrated in Table B.4. Then an initial total sample size N is calculated based on the minimal clinical relevant value $\delta = 0.8$, $\tau^2 = 0.02$, $\#t = 15$ and expected probability of event in the control group $p_C = 0.1$. The initial sample size is 8606. Given the total sample size, we will have a trial size of 574 as 15 is the initial number of trials as defined in Table B.4.

Sequential meta-analysis simulation

The number of events in the two treatment arms are simulated for the 15 trials, where each treatment arm has a total of 287 participants. Analysing the first trial and calculating the z -score to -0.14, we have an inconclusive trial on a 0.05 significance level. This means we will continue to add a trial until a stopping criteria is met.

# Trials	z -score	Boundary	AIS	RIS	$\hat{\tau}^2$
2	-1.06	-20	1148	9391	0
3	-1.98	-4.95	1722	8842	0
4	-2.48	-4.25	2296	8842	0
5	-2.2	-3.81	2870	8985	0
6	-1.93	-3.46	3444	8966	0
7	-1.56	-3.24	4018	9218	0
8	-1.39	-2.99	4592	9055	0
9	-1.91	-4.9	5166	25917	0.0203
10	-2.08	-20	5740	49943	0.0101
11	-1.55	-4.19	6314	23436	0.0238
12	-1.78	-3.33	6888	16421	0.0169
13	-2.02	-2.9	7462	13632	0.0108
14	-2.16	-2.34	8036	9803	0.0043
15	-1.87	-2.31	8610	10230	0.0073
16	-2.2	-2.48	9184	12432	0.0157
17	-2.5	-2.47	9758	13028	0.0187

Table B.5: Outcome of each meta-analysis from one simulation of a retrospective sequential meta-analysis.

How the simulation turned out is shown in Table B.5. Different from the prospective design is the RIS calculated a new each time a new trial is added. This is why the RIS is not constant and also the reason why the boundaries are not always going towards 0 as the number of trials increases. The simulation stops at trial 17 where it crosses the boundary. We further see that this specific simulation ends with an estimate close to the true value of the heterogeneity per Table B.4.

B.4 Results

We are now ready to present some of the results from the simulation study. The simulations were designed with a wanted nominal level of type-I-error of 5% and a type-II-error of 10%. The type-I-errors will be presented for prospective meta-analyses in Table B.6 Panel A and Table B.7 Panel A, where the first table is for simulations simulated with no heterogeneity and the last table is for simulations simulated with heterogeneity. Table B.6 Panel B and Table B.7 Panel B expresses the type-II-error for prospective sequential meta-analyses.

A similar pattern is used for retrospective sequential meta-analyses. The type-I-errors will be presented for retrospective meta-analyses in Table B.8 Panel A and Table B.9 Panel A, where the first table is for simulations simulated with no heterogeneity and the last table is for simulations simulated with heterogeneity. Table B.8 Panel B and Table B.9 Panel B expresses the type-II-error for retrospective sequential meta-analyses.

Prospective sequential meta-analysis

One scenario is left out in the tables below for the prospective meta-analyses. When we have an expected heterogeneity of 0.05, five trials are too few to have a well-powered random-effects meta-analysis. Hence results for this scenario will not be provided.

Note that the sample size calculation for prospective sequential meta-analyses will be adjusted by the sequential design. For this reason is the sample size calculation called "SMA required sample size", where SMA stands for sequential meta-analysis.

A) Prospective SMA	$\theta = 1$ and $\tau^2 = 0$	
Number of studies	#t = 5	#t = 15
Type-I-error	5%	5%
SMA required sample size	8805	8957
Average simulated sample size	8805	8897
Average final number of trials	5	14.9

B) Prospective SMA	$\theta = 0.8$ and $\tau^2 = 0$	
Number of studies	#t = 5	#t = 15
Type-II-error	10%	10%
SMA required sample size	8805	8957
Average simulated sample size	6368	6109
Average final number of trials	3.7	10.3

Table B.6: Empirical type-I- and type-II-error rates based on 10000 simulations per scenario with no heterogeneity for prospective sequential meta-analyses. SMA stands for sequential meta-analysis.

A) Prospective SMA	$\theta = 1$			
	$\tau^2 = 0.02$		$\tau^2 = 0.05$	
	#t = 5	#t = 15	#t = 5*	#t = 15
Number of studies				
Type-I-error	28%	7%	-	10%
SMA required sample size	56776	12536	-	30364
Average simulated sample size	49963	12369	-	29352
Average final number of trials	4.4	14.8	-	14.5

B) Prospective SMA	$\theta = 0.8$			
	$\tau^2 = 0.02$		$\tau^2 = 0.05$	
	#t = 5	#t = 15	#t = 5*	#t = 15
Number of initial studies*				
Type-II-error	9%	13%	-	13%
Required sample size	56776	12536	-	30364
Average simulated sample size	30659	8775	-	19838
Average final number of trials	2.7	10.5	-	9.8

Table B.7: Empirical type-I- and type-II-error rates based on 10000 simulations per scenario with low to modest heterogeneity. The number of studies is a way to control the trial size. A large number of initial trials translates to smaller individual trials. *It is not possible to calculate the sample size with only 5 trials and $\tau^2 = 0.05$. SMA stands for sequential meta-analysis.

Retrospective sequential meta-analysis

One scenario is left out in the tables below for the retrospective meta-analyses. When we have heterogeneity of 0.02 and consider five trials the first analysis of the first study will almost always be significant. Hence most of the results are concerning the analysis of a single trial. We are interested in the sequential behaviour and will for this reason not include the scenario of 0.02 and five trials for retrospective sequential meta-analysis.

A) Retrospective SMA	$\theta = 1$ and $\tau^2 = 0$	
Number of initial studies*	#t = 5	#t = 15
Type-I-error	4%	5%
Required information size	8606	8606
Average simulated information size	10710	9696
Average final number of trials	6.2	16.9

B) Retrospective SMA	$\theta = 0.8$ and $\tau^2 = 0$	
Number of initial studies*	#t = 5	#t = 15
Type-II-error	7%	8%
Required sample size	8606	8606
Average simulated sample size	7918	8147
Average final number of trials	4.6	12.4

Table B.8: Empirical type-I- and type-II-error rates based on 10000 simulations per scenario with no heterogeneity. The number of initial studies is a way to control the trial size. For a larger number of initial trials, the smaller the individual trial size. As the simulations are retrospective the final number of trials may be between 1 and 50. SMA stands for sequential meta-analysis.

B.5 Preliminary conclusions

For prospective sequential meta-analysis when there is no simulated heterogeneity, there will be complete control of the type-I- and type-II-error using the RTSA version of TSA hitting the nominal levels of respectively 5% and 10%. In this scenario, it is indifferent whether few (5) or many (15) studies was part of the the sequential meta-analysis.

For prospective sequential meta-analysis it is not possible to control the nominal levels of type-I- and type-II-errors when there are heterogeneity. In both scenarios of low to moderate heterogeneity, we find that the levels of type-I- and type-II-errors are above the nominal level. Decreasing the size of the individual trials used will have a positive effect on the empirical rates as they converge towards the nominal levels.

Surprisingly it is seen from the retrospective sequential meta-analysis that we have almost complete control of the type-I-errors when the meta-analyses are simulated with no heterogeneity. The type-II-errors will in this scenario be a little bias towards greater power than the nominal level. Similarly to the prospective meta-analysis, we find that heterogeneity will challenge the control of the type-I-errors there we again find better results when heterogeneity is low and one uses many smaller trials compared to fewer larger trials.

A) Retrospective SMA		$\theta = 1$			
		$\tau^2 = 0.02$		$\tau^2 = 0.05$	
Number of initial studies*		#t = 5	#t = 15	#t = 5	#t = 15
Type-I-error	-		13%	59%	42%
Required sample size	-		12045	35126**	29175
Average simulated sample size	-		11563	24588	21784
Average final number of trials	-		14.4	3.5	11.2

B) Retrospective SMA		$\theta = 0.8$			
		$\tau^2 = 0.02$		$\tau^2 = 0.05$	
Number of initial studies*		#t = 5	#t = 15	#t = 5	#t = 15
Type-II-error	-		11%	13%	13%
Required sample size	-		12045	35126**	29175
Average simulated sample size	-		8994	17563	16533
Average final number of trials	-		11.2	2.5	8.5

Table B.9: Empirical type-I- and type-II-error rates based on 10000 simulations per scenario with small to modest heterogeneity. The number of initial studies is a way to control the trial size. *A large number of initial trials translates to smaller individual trials. As the simulations are retrospective the final number of trials may be between 1 and 50. **This required sample size calculation is based on 14 trials as it was not possible to calculate the required sample size with 5 trials. SMA stands for sequential meta-analysis.

Bibliography

- Adelstein, B. A. et al. (2011). “A systematic review and meta-analysis of KRAS status as the determinant of response to anti-EGFR antibodies and the impact of partner chemotherapy in metastatic colorectal cancer”. In: *European Journal of Cancer* 47.9, pp. 1343–1354. ISSN: 0959-8049. DOI: <https://doi.org/10.1016/j.ejca.2011.03.031>. URL: <https://www.sciencedirect.com/science/article/pii/S0959804911002450>.
- Anderson, K. (2022). *gsDesign: Group Sequential Design*. R package version 3.3.0. URL: <https://CRAN.R-project.org/package=gsDesign>.
- Armitage, P. (1957). “Restricted Sequential Procedures”. In: *Biometrika* 44.1-2, pp. 9–26. DOI: 10.1093/biomet/44.1-2.9. URL: <http://dx.doi.org/10.1093/biomet/44.1-2.9>.
- Armitage, P., C. K. McPherson, and B. C. Rowe (1969). “Repeated Significance Tests on Accumulating Data”. In: *Journal of the Royal Statistical Society, Series A (General)* 132.2, p. 235. DOI: 10.2307/2343787. URL: <https://doi.org/10.2307/2343787>.
- Barnard, G. A. (1946). “Sequential Tests in Industrial Statistics”. In: *Supplement to the Journal of the Royal Statistical Society* 8.1, p. 1. DOI: 10.2307/2983610. URL: <http://dx.doi.org/10.2307/2983610>.
- Biggerstaff, B. J. and R. L. Tweedie (1997). “Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis”. In: *Statistics in Medicine* 16.7, pp. 753–768.
- Borenstein, M. et al. (2009). *Introduction to Meta-Analysis*. Wiley.
- Chalmers, I. et al. (2014). “How to increase value and reduce waste when research priorities are set”. In: *The Lancet* 383.9912, pp. 156–165. ISSN: 0140-6736. DOI: [https://doi.org/10.1016/S0140-6736\(13\)62229-1](https://doi.org/10.1016/S0140-6736(13)62229-1). URL: <https://www.sciencedirect.com/science/article/pii/S0140673613622291>.
- Cheung, M. W. L. (2008). “A Model for Integrating Fixed-, Random-, and Mixed-Effects Meta-Analyses Into Structural Equation Modeling.” In: *Psychological Methods* 13.3, pp. 182–202. DOI: 10.1037/a0013163. URL: <https://doi.org/10.1037/a0013163>.
- Deeks, J. J., J. P. T. Higgins, and D. G. Altman (2019). “Analysing data and undertaking meta-analyses”. In: *Cochrane Handbook for Systematic Reviews of Interventions version 6.0*. Chap. 10.
- Demets, D. L. and K. K. G. Lan (1994). “Interim Analysis: the Alpha Spending Function Approach”. In: *Statistics in Medicine* 13.13-14, pp. 1341–1352. DOI: 10.1002/sim.4780131308. URL: <https://doi.org/10.1002/sim.4780131308>.

- Denne, J. S. (2000). "Estimation Following Extension of a Study on the Basis of Conditional Power". In: *Journal of Biopharmaceutical Statistics* 10.2, pp. 131–144. DOI: 10.1081/bip-100101018. URL: <https://doi.org/10.1081/bip-100101018>.
- DerSimonian, R. and N. Laird (1986). "Meta-Analysis in Clinical Trials". In: *Controlled Clinical Trials* 7.3, pp. 177–188. DOI: 10.1016/0197-2456(86)90046-2. URL: [http://dx.doi.org/10.1016/0197-2456\(86\)90046-2](http://dx.doi.org/10.1016/0197-2456(86)90046-2).
- Duhailib, Z. A. et al. (2024). "jscpjgradej/scpj Pearls and Pitfalls-Part 1: Systematic Reviews and Meta-analyses". In: *Acta Anaesthesiologica Scandinavica* nil.nil, nil. DOI: 10.1111/aas.14386. URL: <http://dx.doi.org/10.1111/aas.14386>.
- Elliott, J. H. et al. (2017). "Living Systematic Review: 1. Introduction-The Why, What, When, and How". In: *Journal of Clinical Epidemiology* 91.nil, pp. 23–30. DOI: 10.1016/j.jclinepi.2017.08.010. URL: <http://dx.doi.org/10.1016/j.jclinepi.2017.08.010>.
- Ellis, S. P. and J. W. Stewart (2009). "Temporal Dependence and Bias in Meta-Analysis". In: *Communications in Statistics - Theory and Methods* 38.15, pp. 2453–2462. DOI: 10.1080/03610920802562772. URL: <https://doi.org/10.1080/03610920802562772>.
- Emerson, S. S. and T. R. Fleming (1989). "Symmetric Group Sequential Test Designs". In: *Biometrics* 45.3, p. 905. DOI: 10.2307/2531692. URL: <http://dx.doi.org/10.2307/2531692>.
- Fan, X. F., D. L. DeMets, and K. K. G. Lan (2004). "Conditional Bias of Point Estimates Following a Group Sequential Test". In: *Journal of Biopharmaceutical Statistics* 14.2, pp. 505–530. DOI: 10.1081/bip-120037195. URL: <https://doi.org/10.1081/bip-120037195>.
- Fearon, P. and P. Langhorne (2012). "Services for reducing duration of hospital care for acute stroke patients". In: *Cochrane Database of Systematic Reviews*, nil. DOI: 10.1002/14651858.cd000443.pub3. URL: <https://doi.org/10.1002/14651858.cd000443.pub3>.
- Fehrenbacher, L. et al. (2016). "Atezolizumab Versus Docetaxel for Patients With Previously Treated Non-Small-Cell Lung Cancer (POPLAR): a Multicentre, Open-Label, Phase 2 Randomised Controlled Trial". In: *The Lancet* 387.10030, pp. 1837–1846. DOI: 10.1016/s0140-6736(16)00587-0. URL: [https://doi.org/10.1016/s0140-6736\(16\)00587-0](https://doi.org/10.1016/s0140-6736(16)00587-0).
- Fehrenbacher, L. et al. (2018). "Updated Efficacy Analysis Including Secondary Population Results for Oak: a Randomized Phase Iii Study of Atezolizumab Versus Docetaxel in Patients With Previously Treated Advanced Non-Small Cell Lung Cancer". In: *Journal of Thoracic Oncology* 13.8, pp. 1156–1170. DOI: 10.1016/j.jtho.2018.04.039. URL: <https://doi.org/10.1016/j.jtho.2018.04.039>.
- Firebaugh, G. (1978). "A Rule for Inferring Individual-Level Relationships from Aggregate Data". In: *American Sociological Review* 43.4, pp. 557–572. ISSN: 00031224. URL: <http://www.jstor.org/stable/2094779> (visited on 09/26/2022).
- Fisher, D. J. et al. (2017). "Meta-analytical methods to identify who benefits most from treatments: daft, deluded, or deft approach?" In: *BMJ* 356, j573. DOI: <http://dx.doi.org/10.1136/bmj.j573>.
- Fisher, D. et al. (2011). "A Critical Review of Methods for the Assessment of Patient-Level Interactions in Individual Participant Data Meta-Analysis of

- Randomized Trials, and Guidance for Practitioners”. In: *Journal of Clinical Epidemiology* 64.9, pp. 949–967. DOI: 10.1016/j.jclinepi.2010.11.016. URL: <https://doi.org/10.1016/j.jclinepi.2010.11.016>.
- Fogarty, M. et al. (2018). “Delayed Vs Early Umbilical Cord Clamping for Preterm Infants: a Systematic Review and Meta-Analysis”. In: *American Journal of Obstetrics and Gynecology* 218.1, pp. 1–18. DOI: 10.1016/j.ajog.2017.10.231. URL: <https://doi.org/10.1016/j.ajog.2017.10.231>.
- Godolphin, P. J. et al. (2022). “Estimating Interactions and Subgroup-specific Treatment Effects in Meta-analysis Without Aggregation Bias: a Within-trial Framework”. In: *Research Synthesis Methods* 14.1, pp. 68–78. DOI: 10.1002/jrsm.1590. URL: <http://dx.doi.org/10.1002/jrsm.1590>.
- Greenland, S. and H. Morgenstern (Mar. 1989). “Ecological Bias, Confounding, and Effect Modification”. In: *International Journal of Epidemiology* 18.1, pp. 269–274. ISSN: 0300-5771. DOI: 10.1093/ije/18.1.269. eprint: <https://academic.oup.com/ije/article-pdf/18/1/269/1702405/18-1-269.pdf>. URL: <https://doi.org/10.1093/ije/18.1.269>.
- group, C. (2021). *CONSORT 2010*. (accessed Sep. 13 2021). CONSORT group. URL: <http://www.consort-statement.org/consort-2010>.
- Harrer, M. et al. (2019). *dmear: Companion R Package For The Guide 'Doing Meta-Analysis in R'*. R package version 0.0.9000. URL: <http://dmetar.protectlab.org/>.
- Higgins, J. P. T. (2003). “Measuring Inconsistency in Meta-Analyses”. In: *BMJ* 327.7414, pp. 557–560. DOI: 10.1136/bmj.327.7414.557. URL: <http://dx.doi.org/10.1136/bmj.327.7414.557>.
- Higgins, J. P. T. and S. G. Thompson (2002). “Quantifying Heterogeneity in a Meta-Analysis”. In: *Statistics in Medicine* 21.11, pp. 1539–1558. DOI: 10.1002/sim.1186. URL: <http://dx.doi.org/10.1002/sim.1186>.
- Higgins, J. P. T., A. Whitehead, and M. Simmonds (2010). “Sequential Methods for Random-Effects Meta-Analysis”. In: *Statistics in Medicine* 30.9, pp. 903–921. DOI: 10.1002/sim.4088. URL: <https://doi.org/10.1002/sim.4088>.
- Hong, H. et al. (2016). “A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons”. In: *Research Synthesis Methods* 7.1, pp. 6–22. DOI: <https://doi.org/10.1002/jrsm.1153>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/jrsm.1153>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/jrsm.1153>.
- Hua, H. et al. (2016). “One-Stage Individual Participant Data Meta-Analysis Models: Estimation of Treatment-Covariate Interactions Must Avoid Ecological Bias By Separating Out Within-Trial and Across-Trial Information”. In: *Statistics in Medicine* 36.5, pp. 772–789. DOI: 10.1002/sim.7171. URL: <https://doi.org/10.1002/sim.7171>.
- Hwang, I. K., W. J. Shih, and J. S. D. Cani (1990). “Group Sequential Designs Using a Family of Type I Error Probability Spending Functions”. In: *Statistics in Medicine* 9.12, pp. 1439–1445. DOI: 10.1002/sim.4780091207. URL: <http://dx.doi.org/10.1002/sim.4780091207>.
- Imberger, G. et al. (2015). “Systematic Reviews of Anesthesiologic Interventions Reported As Statistically Significant”. In: *Anesthesia & Analgesia* 121.6, pp. 1611–1622. DOI: 10.1213/ane.0000000000000892. URL: <http://dx.doi.org/10.1213/ANE.0000000000000892>.

- Imberger, G. et al. (2016). “False-Positive Findings in Cochrane Meta-Analyses With and Without Application of Trial Sequential Analysis: an Empirical Review”. In: *BMJ Open* 6.8, e011890. DOI: 10.1136/bmjopen-2016-011890. URL: <http://dx.doi.org/10.1136/bmjopen-2016-011890>.
- IntHout, J., J. P. Ioannidis, and G. F. Borm (2014). “The Hartung-Knapp-Sidik-Jonkman Method for Random Effects Meta-Analysis Is Straightforward and Considerably Outperforms the Standard Dersimonian-Laird Method”. In: *BMC Medical Research Methodology* 14.1, p. 25. DOI: 10.1186/1471-2288-14-25. URL: <https://doi.org/10.1186/1471-2288-14-25>.
- Ioannidis, J. P. A. (2022). “Correction: Why Most Published Research Findings Are False”. In: *PLOS Medicine* 19.8, e1004085. DOI: 10.1371/journal.pmed.1004085. URL: <http://dx.doi.org/10.1371/journal.pmed.1004085>.
- Jackson, D. and R. Turner (2017). “Power Analysis for Random-Effects Meta-Analysis”. In: *Research Synthesis Methods* 8.3, pp. 290–302. DOI: 10.1002/jrsm.1240. URL: <http://dx.doi.org/10.1002/jrsm.1240>.
- Jennison, C. and B. W. Turnbull (1984). “Repeated Confidence Intervals for Group Sequential Clinical Trials”. In: *Controlled Clinical Trials* 5.1, pp. 33–45. DOI: 10.1016/0197-2456(84)90148-x. URL: [http://dx.doi.org/10.1016/0197-2456\(84\)90148-x](http://dx.doi.org/10.1016/0197-2456(84)90148-x).
- Jennison, C. and B. Turnbull (Sept. 1999). *Group sequential tests with applications to clinical trials*. English. Chapman and Hall/CRC Interdisciplinary Statistics. UK United Kingdom: Chapman and Hall. ISBN: 9780849303166.
- Kleinert, S. et al. (2014). *Further emphasis on research in context*. URL: <https://www.thelancet.com/action/showPdf?pii=S0140-6736%2814%2962047-X>.
- Kulinskaya, E., R. Huggins, and S. H. Dogo (2015). “Sequential Biases in Accumulating Evidence”. In: *Research Synthesis Methods* 7.3, pp. 294–305. DOI: 10.1002/jrsm.1185. URL: <https://doi.org/10.1002/jrsm.1185>.
- Kulinskaya, E. and J. Wood (2013). “Trial Sequential Methods for Meta-Analysis”. In: *Research Synthesis Methods* 5.3, pp. 212–220. DOI: 10.1002/jrsm.1104. URL: <https://doi.org/10.1002/jrsm.1104>.
- Lan, K. K. G. and D. L. DeMets (1983). “Discrete Sequential Boundaries for Clinical Trials”. In: *Biometrika* 70.3, p. 659. DOI: 10.2307/2336502. URL: <https://doi.org/10.2307/2336502>.
- Lee, K. W. C. et al. (2019). “The Impact of Smoking on the Effectiveness of Immune Checkpoint Inhibitors - a Systematic Review and Meta-Analysis”. In: *Acta Oncologica* 59.1, pp. 96–100. DOI: 10.1080/0284186x.2019.1670354. URL: <https://doi.org/10.1080/0284186x.2019.1670354>.
- Liu, A. et al. (2004). “Conditional Maximum Likelihood Estimation Following a Group Sequential Test”. In: *Biometrical Journal* 46.6, pp. 760–768. DOI: 10.1002/bimj.200410076. URL: <https://doi.org/10.1002/bimj.200410076>.
- Marschner, I. C., M. Schou, and A. J. Martin (2022). “Estimation of the Treatment Effect Following a Clinical Trial That Stopped Early for Benefit”. In: *Statistical Methods in Medical Research* 31.12, pp. 2456–2469. DOI: 10.1177/09622802221122445. URL: <http://dx.doi.org/10.1177/09622802221122445>.
- Meta-Analysis with R* (2015). Springer International Publishing Switzerland.

- Pampallona, S. and A. A. Tsiatis (1994). “Group Sequential Designs for One-Sided and Two-Sided Hypothesis Testing With Provision for Early Stopping in Favor of the Null Hypothesis”. In: *Journal of Statistical Planning and Inference* 42.1-2, pp. 19–35. DOI: 10.1016/0378-3758(94)90187-2. URL: [http://dx.doi.org/10.1016/0378-3758\(94\)90187-2](http://dx.doi.org/10.1016/0378-3758(94)90187-2).
- Pinheiro, J. et al. (2020). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-151. URL: <https://CRAN.R-project.org/package=nlme>.
- Pogue, J. M. and S. Yusuf (1997). “Cumulating Evidence From Randomized Trials: Utilizing Sequential Monitoring Boundaries for Cumulative Meta-Analysis”. In: *Controlled Clinical Trials* 18.6, pp. 580–593. DOI: 10.1016/S0197-2456(97)00051-2. URL: [https://doi.org/10.1016/S0197-2456\(97\)00051-2](https://doi.org/10.1016/S0197-2456(97)00051-2).
- R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/>.
- Rabe, H. et al. (2012). “Effect of timing of umbilical cord clamping and other strategies to influence placental transfusion at preterm birth on maternal and infant outcomes”. In: *Cochrane Database of Systematic Reviews* 8. ISSN: 1465-1858. DOI: 10.1002/14651858.CD003248.pub3. URL: <https://doi.org/10.1002/14651858.CD003248.pub3>.
- Rabe, H. et al. (2019). “Effect of Timing of Umbilical Cord Clamping and Other Strategies To Influence Placental Transfusion At Preterm Birth on Maternal and Infant Outcomes”. In: *Cochrane Database of Systematic Reviews* 2019.9, nil. DOI: 10.1002/14651858.cd003248.pub4. URL: <http://dx.doi.org/10.1002/14651858.cd003248.pub4>.
- Riley, R. D. et al. (2020). “Individual Participant Data Meta-analysis To Examine Interactions Between Treatment Effect and Participant-level Covariates: Statistical Recommendations for Conduct and Planning”. In: *Statistics in Medicine* 39.15, pp. 2115–2137. DOI: 10.1002/sim.8516. URL: <https://doi.org/10.1002/sim.8516>.
- SAS Institute Inc. (2003). *SAS Software, Version 9.4*. Cary, NC. URL: <http://www.sas.com/>.
- Seide, S. E., C. Rover, and T. Friede (2019). “Likelihood-Based Random-Effects Meta-Analysis With Few Studies: Empirical and Simulation Studies”. In: *BMC Medical Research Methodology* 19.1, p. 16. DOI: 10.1186/s12874-018-0618-3. URL: <http://dx.doi.org/10.1186/s12874-018-0618-3>.
- Seidler, A. L. et al. (2019). “A Guide To Prospective Meta-Analysis”. In: *BMJ* nil.nil, p. 15342. DOI: 10.1136/bmj.15342. URL: <http://dx.doi.org/10.1136/bmj.15342>.
- Simmonds, M., G. Salanti, and J. McKenzie (2017). “Living Systematic Reviews: 3. Statistical Methods for Updating Meta-Analyses”. In: *Journal of Clinical Epidemiology* 91.nil, pp. 38–46. DOI: 10.1016/j.jclinepi.2017.08.008. URL: <http://dx.doi.org/10.1016/j.jclinepi.2017.08.008>.
- Soerensen, A. L. and I. C. Marschner (2023a). “Linear Mixed Models for Investigating Effect Modification in Subgroup Meta-Analysis”. In: *Statistical Methods in Medical Research* 32.5, pp. 994–1009. DOI: 10.1177/09622802231163330. URL: <http://dx.doi.org/10.1177/09622802231163330>.

- Soerensen, A. L. and M. H. Olsen (2023b). *RTSA: 'Trial Sequential Analysis' for Error Control and Inference in Sequential Meta-Analyses*. R package version 0.2.1. URL: <https://github.com/AnneLyng/RTSA>.
- Spence, G. T., D. Steinsaltz, and T. R. Fanshawe (2016). “A Bayesian Approach To Sequential Meta-analysis”. In: *Statistics in Medicine* 35.29, pp. 5356–5375. DOI: 10.1002/sim.7052. URL: <https://doi.org/10.1002/sim.7052>.
- Strickland, P. A. O. and G. Casella (2003). “Conditional Inference Following Group Sequential Testing”. In: *Biometrical Journal* 45.5, pp. 515–526. DOI: 10.1002/bimj.200390029. URL: <http://dx.doi.org/10.1002/bimj.200390029>.
- Sutton, A. J. et al. (2000). *Methods for meta-analysis in medical research*. Vol. 348. Wiley Chichester.
- Tarnow-Mordi, W. et al. (2017). “Delayed Versus Immediate Cord Clamping in Preterm Infants”. In: *New England Journal of Medicine* 377.25, pp. 2445–2455. DOI: 10.1056/nejmoa1711281. URL: <https://doi.org/10.1056/nejmoa1711281>.
- Tarnow-Mordi, W. O. et al. (2014). “Timing of cord clamping in very preterm infants: more evidence is needed”. In: *American Journal of Obstetrics and Gynecology* 211.2, pp. 118–123. ISSN: 0002-9378. DOI: <https://doi.org/10.1016/j.ajog.2014.03.055>. URL: <https://www.sciencedirect.com/science/article/pii/S000293781400283X>.
- Tarnow-Mordi, W. O. et al. (2020). “The Effect of Lactoferrin Supplementation on Death Or Major Morbidity in Very Low Birthweight Infants (LIFT): a Multicentre, Double-Blind, Randomised Controlled Trial”. In: *The Lancet Child & Adolescent Health* 4.6, pp. 444–454. DOI: 10.1016/s2352-4642(20)30093-6. URL: [https://doi.org/10.1016/s2352-4642\(20\)30093-6](https://doi.org/10.1016/s2352-4642(20)30093-6).
- ter Schure, J. and P. Grünwald (2019). “Accumulation Bias in meta-analysis: the need to consider time in error control”. In: *F1000Research* 8, p. 962. ISSN: 2046-1402. DOI: 10.12688/f1000research.19375.1. URL: <http://dx.doi.org/10.12688/f1000research.19375.1>.
- Thomas, J. et al. (2023). *Cochrane Handbook for Systematic Reviews of Interventions*. version 6.4 (updated August 2023). Chap. 22: Prospective approaches to accumulating evidence. URL: www.training.cochrane.org/handbook.
- Thorlund, K. et al. (2011). *User manual for trial sequential analysis (TSA)*. Copenhagen Trial Unit, Centre for Clinical Intervention Research. Copenhagen, Denmark. URL: https://ctu.dk/wp-content/uploads/2021/03/2017-10-10-TSA-Manual-ENG_ER.pdf.
- Turner, R. M. et al. (2014). “Predictive Distributions for Between-Study Heterogeneity and Simple Methods for Their Application in Bayesian Meta-Analysis”. In: *Statistics in Medicine* 34.6, pp. 984–998. DOI: 10.1002/sim.6381. URL: <https://doi.org/10.1002/sim.6381>.
- Viechtbauer, W. (2006). “Confidence Intervals for the Amount of Heterogeneity in Meta-Analysis”. In: *Statistics in Medicine* 26.1, pp. 37–52. DOI: 10.1002/sim.2514. URL: <http://dx.doi.org/10.1002/sim.2514>.
- (2010). “Conducting meta-analyses in R with the metafor package”. In: *Journal of Statistical Software* 36.3, pp. 1–48. URL: <https://www.jstatsoft.org/v36/i03/>.
- Wald, A. (1947). *Sequential analysis*. John Wiley.

- Wang, S. K. and A. A. Tsiatis (1987). “Approximately Optimal One-Parameter Boundaries for Group Sequential Trials”. In: *Biometrics* 43.1, p. 193. DOI: 10.2307/2531959. URL: <http://dx.doi.org/10.2307/2531959>.
- Wassmer, G. and W. Brannath (2016). *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Springer Series in Pharmaceutical Statistics. Springer International Publishing, nil. DOI: 10.1007/978-3-319-32562-0. URL: <https://doi.org/10.1007/978-3-319-32562-0>.
- Wassmer, G. and F. Pahlke (2022). *rpact: Confirmatory Adaptive Clinical Trial Design and Analysis*. R package version 3.3.1. URL: <https://CRAN.R-project.org/package=rpact>.
- Wetterslev, J., J. C. Jakobsen, and C. Gluud (2017). “Trial Sequential Analysis in Systematic Reviews With Meta-Analysis”. In: *BMC Medical Research Methodology* 17.1, p. 39. DOI: 10.1186/s12874-017-0315-7. URL: <http://dx.doi.org/10.1186/s12874-017-0315-7>.
- Wetterslev, J. et al. (2008). “Trial Sequential Analysis May Establish When Firm Evidence Is Reached in Cumulative Meta-Analysis”. In: *Journal of Clinical Epidemiology* 61.1, pp. 64–75. DOI: 10.1016/j.jclinepi.2007.03.013. URL: <https://doi.org/10.1016/j.jclinepi.2007.03.013>.
- (2009). “Estimating Required Information Size By Quantifying Diversity in Random-Effects Model Meta-Analyses”. In: *BMC Medical Research Methodology* 9.1, p. 86. DOI: 10.1186/1471-2288-9-86. URL: <https://doi.org/10.1186/1471-2288-9-86>.
- Wetterslev, J. et al. (2015). “The Effects of High Perioperative Inspiratory Oxygen Fraction for Adult Surgical Patients”. In: *Cochrane Database of Systematic Reviews* 2016.9, nil. DOI: 10.1002/14651858.cd008884.pub2. URL: <http://dx.doi.org/10.1002/14651858.CD008884.pub2>.
- Whitehead, A. (2002). *Meta-analysis of Controlled Clinical Trials*. 2002 John Wiley & Sons, Ltd.