Some Statistical Problems in Sequential Meta-analysis

April 20, 2016

Some Statistical Problems in Sequential Meta-analysis

Samson Henry Dogo

Supervisors

Prof. Elena Kulinskaya and Dr. Allan Clark

A thesis submitted in fulfilment of the requirements

for the degree of Doctor of Philosophy

at the University of East Anglia,

School of Computing Sciences, Norwich, April 2016.

Abstract

The objective of meta-analysis is to combine results from several independent studies in order to make evidence more generalisable and provide evidence base for decision making. However, recent studies show that the magnitude of effect size estimates reported in many areas of research have significantly changed over time. These temporal trends can be dramatic and even lead to the loss or gain of the statistical significance of the cumulative treatment effect (Kulinskaya and Koricheva, 2010). Standard sequential methods including cumulative meta-analysis, sequential meta-analysis, the use of quality control charts and penalised z-test have been proposed for monitoring the trends in meta-analysis. But these methods are only effective when monitoring in fixed effect model (FEM) of meta-analysis. For random-effects model (REM), the analysis incorporates the heterogeneity variance, τ^2 and its estimation creates complications. This thesis proposes the use of a truncated CUSUM-type test (Gombay method) for sequential monitoring in REM, and also examines the effect of accumulating evidence in meta-analysis. Simulations show that the use of Gombay method with critical values derived from asymptotic theory does not control the Type I error. However, the test with bootstrap-based critical values (retrospective Gombay sequential bootstrap test for REM) leads to a reduction of the difference between the true and nominal levels, and thus constitutes a good approach for monitoring REM. Application of the proposed method is illustrated using two meta-analytic examples from medicine. Two kinds of bias associated with accumulating evidence, termed “sequential decision bias” and “sequential design bias” are identified. It was demonstrated analytically and by simulations that both types of sequential biases are non negligible. Simulations also show that sequential biases increase with increased heterogeneity.

Contents

List of Figures

List of Tables

Acknowledgements

Glory be to God almighty for giving me the opportunity to reach this feet and complete my studies successfully. May his name be praised forever and ever. Amen.

I wish to express my profound gratitude and sincere thanks to my first supervisor professor Elena Kulinskaya, for giving me the opportunity to do my PhD at the University of East Anglia, for teaching me the ABC of meta-analysis, for her constructive criticisms, guidance and encouragement. The patience she exercised and the untiring help in putting me through in all the statistical rigours have been a source of inspiration throughout the stages of my research. I am also grateful to my second supervisor Dr Allan Clark for all the time spent on reading through my work and the valuable contributions. It has been a honour and pleasure for me to work under the guidance of my supervisors.

I would like to thank my research team mates Skevi, Hussain, Ilyas and Lisanne for all their encouragement and assistance during my studies.

Next, I am grateful to my wife, Murna, who stood by my side in every way throughout the period of my studies. She witnessed my sadness and frustration when things were not working and my happiness when things were working. I am particularly grateful for her steadfastness, humility and loving kindness, and I am happy that she is now happier than I am to see this project completed. I am indebted to my late parent, brothers and sisters who guided me right from my childhood to this stage, and can now be proud of having a son that is a PhD holder. Many thanks go to my friends and well wishers for their prayers, moral support and encouragement. To

all others from whom I have drawn inspiration, especially my previous lecturers from former schools who stirred me to work hard, I am very grateful.

Dedication

To Murna, Shallomith and Christina.

Chapter 1

Introduction

1.1 Brief history of meta-analysis

Since the middle of the 20th century, there has been considerable increase in the volume of scientific research in nearly every field with new findings daily challenging the existing evidence. There is a need to carefully summarize the available literature and perform a review of the data. Traditional method of assimilating accumulating information based on discursive reviews can not adequately provide accurate, reliable and valid summaries of research (Glass et al., 1984), and thus more objective methods are required. Meta-analysis is a statistical method that provides the first step to such objectivity (Schmidt, 1992), allows to combine results from many studies and accurately estimate the effect of interest (Hedges, 1987, Rosenthal, 1978). Such analyses have become a very commonly used methodology for quantitative review in the medical and social sciences.

Meta-analysis started with a paper on a medical problem by Pearson (1904). He analysed data on the correlation between inoculation and mortality for different groups of soldiers across the British Empire and found a statistical significance in the effect (correlation coefficient). This is considered to be the first meta-analysis. Pearson was very critical about the consistency of individual study results and how future research can be improved, and thus his work possesses the characteristics of a correct meta-analysis (Vorosbcsuk, 2010). Further contributions and advances in the subject were made by Cochran (1937), Fisher (1934), Pearson (1933), Tippett et al. (1931), Yates and Cochran (1938). In particular, Pearson (1933) and Tippett et al. (1931) independently proposed a method for combining statistical tests using the product of the p-values across studies. Pearson (1933) commented that "*when a number of independent tests of significance have been made, it sometimes happens that although few or none are significant, the aggregate gives an impression that the probabilities are on the whole lower than would often have been obtained by chance*". These early procedures for statistically combining results of independent studies, though important, were under utilized (Cooper, 2007).

The use of meta-analysis in the social sciences and education research started in the 1970's, first with publications by Glass (1976) and Smith and Glass (1977) on integrating findings in education and psychotherapy. Glass (1976) coined the term "meta-analysis" and defined it as a statistical analysis of large collection of results from individual studies for the purpose of integrating findings. Meta-analysis is a quantitative statistical analysis of several independent studies on the same topic with the purpose of testing the pooled data for statistical significance. Glass (1976) argued that such analyses are needed to contain and make sense of the large volumes of research literature available in education and social sciences. Meta-analysis has gained acceptance and is now used in many areas of science as the formal statistical method for quantitative evidence synthesis. It is used in numerous applications to synthesize and strengthen evidence about the treatment efficacy and provide evidence for decision making. Meta-analysis helps to decide when evidence of benefit or harm of a new intervention is statistically significant and scientifically convincing to adopt or reject the investigated treatment (DerSimonian and Laird, 1986, Leimu and Koricheva, 2004, Pogue and Yusuf, 1997, Kuppens and Onghena, 2012). It is now accepted in medicine as the standard statistical technique used for gauging sufficiency in accumulated evidence. This is evidenced by the rising number of publications using meta-analysis in medical science which has increased exponentially in recent years. For example, the number of publications on MEDLINE about meta-analysis has increased from less than 300 in 1985 to more than 3000 in 2005, see Khoshdel et al. (2006), Kulinskaya and Morgenthaler (2012). In addition, there is large number of books written and published primarily focusing on meta-analytic methods. See Chalmers et al. (2002), Hedges (1987), Hedges and Olkin (1985), Schmidt (1992), Rosenthal and DiMatteo (2001) for further information on the history and recent developments in meta-analysis. Introduction to meta-analysis is provided in Chapter 2.

1.2 Problem of temporal trends in meta-analysis

Meta-analysis is in several ways is a very powerful method of analysis (Arnqvist and Wooster, 1995). It allows one to go beyond the limits of a single study and establish what are the consistent findings about an intervention effect. Meta-analysis makes use of both published and unpublished results, and without it useful information are left fallow or are at least under-utilized. By combining information from several studies meta-analysis allows the combined sample size to achieve a higher statistical power for the outcome of interest compared to the less precise measures derived from single individual studies. The precision with which treatment effect is estimated largely depends on the sample size, and since meta-analysis has larger combined sample size it provides more accurate estimates of the effect of interest. Meta-analysis facilitates the investigation of heterogeneity- a measure of inconsistency of treatment effects across all studies, allows inference on summary estimates and generalisation of evidence. By its ability to extract clear answers from the research literature, it has made a difference in the lives of many patients by providing answers to clinical questions about their care, answers that might not have been obtained from individual studies (Rosenthal and DiMatteo, 2001). Meta-analysis is also used to decide whether enough evidence has been gathered so that further trials are unnecessary.

However recent publications in many areas of research reveal that scientific evidence is not static and tends to change over time. New studies either strengthen or challenge the conclusions of previous findings, resulting in changes in the effects and their variance over time. For example, Hodgson et al. (1989) found a significant decline in the sensitivity of chest X-rays in detecting hypersensitivity pneumonitis of about 1.4 % per annum, which they claimed to be a result of secular trends in knowledge and earlier diagnosis or changes in the disease itself. Nieuwkamp et al. (2009) found a decrease in case fatality of aneurysmal sub-arachnoid haemorrhage during the period 1960-1995, which they attributed to improvement in early diagnostic and treatment strategies. Similar temporal changes have also been reported in education (Hyde et al., 1990), medicine (Gehr et al., 2006), psychology (Brugger et al., 2011, Twenge. et al., 2008, Grabe et al., 2008) to mention but a few. These temporal trends in effect size estimates can be dramatic and often lead to the loss or gain of the statistical significance (Kulinskaya and Koricheva, 2010). If meta-analysis is conducted by ignoring temporal trends when trends are actually present, its results and conclusions can be impaired and any statistical inference about the treatment effect will be misleading. Therefore appropriate statistical techniques that are suitable for monitoring the trends in changes in effect size estimates are required so that results and conclusions of meta-analysis can be interpreted based on the time it was conducted.

A number of sequential methods have been proposed for monitoring the trends in changes in effect size estimates in meta-analysis, see Lau et al. (1992), Leimu and Koricheva (2004), Pogue and Yusuf (1997), Wetterslev et al. (2008), Higgins et al. (2011), Whitehead (1997b), Bollen et al. (2006), Kulinskaya and Koricheva (2010), Lan et al. (2003). The methods allow researchers to gauge sufficiency of evidence (Lau et al., 1992, Pogue and Yusuf, 1997, Wetterslev et al., 2008) and can be used for monitoring the trends in effect size estimates (Leimu and Koricheva, 2004, Kulinskaya and Koricheva, 2010, Ioannidis and Trikalinos, 2005). However these methods of monitoring effect size estimates are based on the solid statistical theory only in the fixed effect model (FEM) of meta-analysis. For random-effects model (REM), the analysis incorporates the heterogeneity variance, τ^2 and its estimation creates complications in the analysis.

Chapter 3 reviews the standard sequential methods in meta-analysis. A new method based on the use of Gombay (2003) truncated CUSUM-type test is proposed in Chapter 4. It is used for sequential change detection for parametric models involving a nuisance parameter. The Gombay method consists of a sequence of score tests about a parameter of interest and terminates at a fixed truncation point, see Chapter 4 for a detailed description of the method. In the application of the Gombay methods in random-effects model of meta-analysis, the heterogeneity parameter, τ^2 is treated as a nuisance parameter, a parameter that is not of immediate interest but must be accounted for in the course of the analysis. The Gombay (2003) method has solid statistical foundations and may constitute a better and more efficient sequential approach to monitoring effect size estimates in random-effects meta-analysis. However, results of simulations given in Chapter 4 show that the test based on the asymptotic critical values suggested in Gombay (2003) is disappointing. Results of this Chapter are published in International Journal of Mathematical, Computational, Statistical, Natural and Physical Engineering, see Dogo et al. (2015). Therefore bootstrap critical values are introduced in Chapter 5 for the use with the Gombay test for sequential random-effects meta-analysis. It is hoped that the new method will provide an alternative approach to sequential random effects meta-analysis as well as stimulate further research on the subject. Results of this Chapter were submitted as Sequential change detection and monitoring of temporal trends in random-effects meta-analysis by Samson Henry Dogo, Allan Clark, Elena Kulinskaya (2015) to Research Synthesis Method for publication.

1.3 The effect of existing evidence on meta-analysis

The idea that results from previous meta-analyses should be used for design of new trials is widely recognised. For example, the UK Medical Research Council requires a comprehensive review of existing evidence before funding trials (Glasziou et al., 2006). The guidelines of several medical journals including the Journal of American Medical Association and the Lancet state that all reports of clinical trials must include a summary with direct reference to existing meta-analyses (Goudie et al., 2010).

There are two ways of using existing evidence to inform further research. The first is using existing information in making decision to conduct a new trial (sequential decision). The second is using previous meta-analyses and systematic reviews to design the next trial (sequential design). That is both the decision to conduct an experiment and the subsequent design of this experiment may depend on the results of previous experiments, and after the new experiment is conducted the results are combined in an updated meta-analysis.

Sequential and cumulative meta-analysis are established statistical methods in fixed and random effects models of meta analyses. See Whitehead (1997a); Higgins et al.

(2011); Bollen et al. (2006); van der Tweel and Bollen (2010) to name a very few. Often, in a sequential analysis after each trial the only decision is whether or not to add the next trial in a sequence of independent trials. Whilst not advocating the approach and remarking on its inherent flaws, van der Tweel and Bollen (2010) noted "*The usual approach is to repeatedly test the null hypothsis of equal effectiveness of two treatments on the cumulative data. If the test result is not statistically significant, a new trial is added and the test is repeated*". Moreover a systematic review can also lead to the conclusion that a new trial is unnecessary (Goudie et al., 2010).

Chapter 6 explores a different approach to standard sequential meta-analysis in that after K studies are accumulated and their results combined, a meta-analyst has an active role in decision-making and the design of subsequent, (K+1)th study, participating in the study team. The effect of evidence from previous meta-analyses on the decision-making and the biases associated with sequential decision and sequential design are examined. Results of this Chapter were accepted for publication as Sequential biases in accumulating evidence by Elena Kulinskaya, Richard Huggins, Samson Henry Dogo in Research Synthesis Methods on the 27-Aug-2015.

1.4 Outline of thesis

The outline of the thesis is as follows. Chapter 2 introduces the preliminaries of meta-analysis including the concept of effect size, its measurement and the models used to combine results from different studies in meta-analysis. Chapter 3 is the introduction to sequential analysis including some sequential designs and review of the methods for monitoring trends in meta-analysis. Chapter 4 introduces a new approach to sequential random-effects meta-analysis. Chapter 5 presents Gombay test for REM with bootstrap critical values. Problems to do with sequential bias in accumulating evidence are discussed in Chapter 6. Chapter 7 is the summary and conclusions of the thesis.

1.5 Publications

- Dogo, S. H., Clark, A., and Kulinskaya, E. (2015). A sequential approach for random-effects meta-analysis. International Journal of Mathematical, Computational, Statistical, Natural and Phisical Engineering, 9(1).

- Dogo, S. H., Clark, A., and Kulinskaya, E. (2015). Sequential change detection and monitoring of temporal trends in random-effects meta-analysis. Submitted on the 15-Oct-2015 for publication in Research Synthesis Methods.

- Kulinskaya, E., Huggins, R., and Dogo, S. H. (2015). Sequential biases in accumulating evidence. Accepted for publication in Research Synthesis Methods on 27-Aug-2015.

Chapter 2

Preliminaries of meta-analysis

This Chapter presents the basics of meta-analysis which are fundamental in understanding the methodologies used in this research. The first Section describes the theoretical concept of effect size and its measurement. The second Section discusses the two models: fixed and random-effects models used to combine results from individual studies in meta-analysis and their statistical properties.

2.1 Theoretical concept of effect size

Traditional methods to establish the presence or otherwise of a treatment effect in a study are often based on the use of p-values, the probability of observing results in the study (or results more extreme) given that the null hypothesis is true. However the p-value is not reliable and has many controversies. The p-value depends on the sample size, see Sullivan and Feinn (2012), Lin et al. (2013). For example, the p-value of a

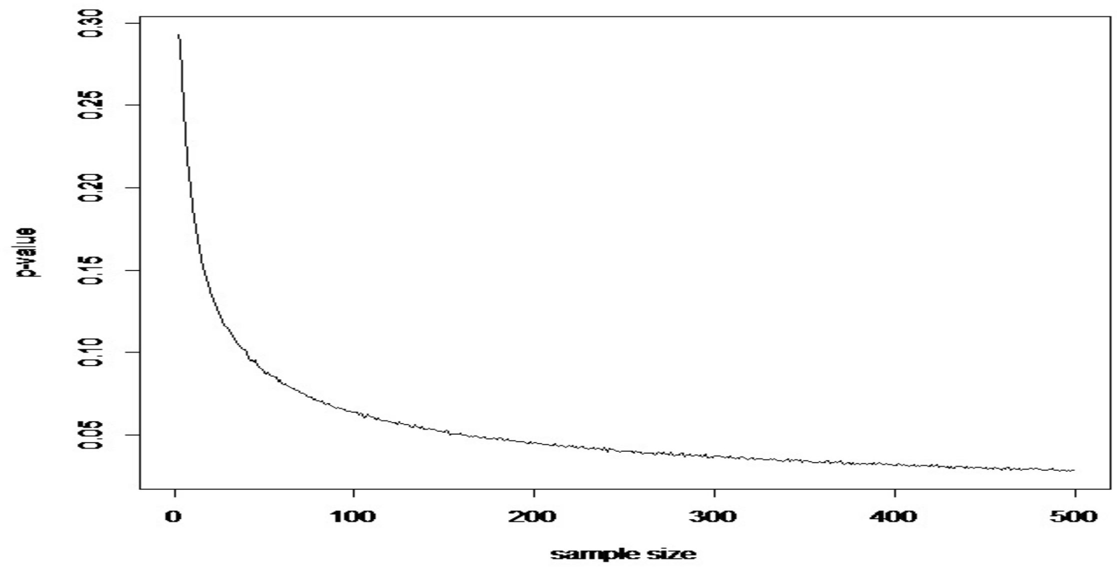


Figure 2.1: Relationship between sample size (n) and the average p-value calculated from data generated from 100000 simulations of $x \sim N(\mu_0, \sigma^2/n)$, $H_0 : \mu_0 = 0$, $\sigma^2 = 0.025$ and sample size n.

two-sided single sample t-test is calculated as

$$\text{p-value} = 2 * tcdf_{n-1}\left\{-\left|\frac{\bar{X}-\mu}{s/\sqrt{n}}\right|\right\} \tag{2.1}$$

where $\bar{X}$ is the sample mean, s is the sample standard deviation, n is the sample size and $tcdf_{n-1}$ is the cumulative t-distribution function with $n-1$ degrees of freedom. Using this formula, a simulation was conducted to explore the relationship between the p-value and the sample size. To do this, data were generated from the normal distribution $x \sim N(\mu_0, \sigma^2/n)$, and the p-value was calculated using (2.1) with $\mu = \mu_0$, $\sigma^2 = 0.025$ and n taking values from 1 to 500. The procedure was repeated 100000 times for each value of n and the average of p-values plotted against n, see Figure 2.1. Clearly, the p-value depends on the sample size n, in fact the p-value tends to zero with increase in the value of n. Therefore, even when there is no treatment effect of practical importance, increase in sample size can lead to a very small p-value and thereby results in false rejection of the null hypothesis (false positive result). The p-value does not inform the researcher of the benefits, harms or magnitude of the treatment effect. If the p-value is small and the null hypothesis is rejected, the researcher can only conclude that the treatment effect is significantly different from zero which has no practical relevance. Anscombe (1956) remarked that the use of p-value is irrelevant, what is needed for researchers is the effect size and its standard error. Effect sizes are a necessary compliment to statistical significance testing because they provide important information that such tests alone can not offer (Ledesma et al., 2009).

Effect size is an alternative statistical tool for evaluating the effect of a treatment. It measures how large or small is a relationship between two or more variables in sampled data. Effect size is the common currency that summaries the findings from a specific area of research (Becker, 2000). It is an objective and standardized measure of the magnitude of the observed effect (Field, 2005). For binary data it is often calculated as odds ratio, risk difference or relative risk. For continuous data it is often calculated as mean difference, means ratio, standardized mean difference or correlation.

Effect sizes are usually presented with their confidence intervals. The confidence interval (CI) is an interval containing the population parameter with a specified level of confidence.

There are many ways to construct the confidence intervals for effect size including the inversion approach (see Venables (1975), Harlow et al. (1997)), bootstrap method (see Efron (1987), Efron and Tibshirani (1994), Efron (1982)), and the most commonly used method which relies on the asymptotic normality of the distribution of effect size (Hedges and Olkin (1985);Hess and Kromrey (2004). The following Sections present the common effect size measures in meta-analysis together with their confidence intervals, based on the asymptotic normality. These CI's at (1-α)% are generally given by

$$\mathrm{CI} = y \pm z_{1-\alpha/2}\sqrt{\mathrm{var}(y)}, \tag{2.2}$$

where y is the effect size measure and $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-th percentile of the standard normal distribution.

2.2 Effect size measures

There exist different types of effect size measures used in meta-analysis depending on the type of data and the objective of the research. Here we consider only those that are relevant to this research.

2.2.1 Effect size measures for continuous data

In meta-analysis, effect size measures for continuous data are used when studies outcomes are measured on a continuous scale. These outcomes include variables such as height, weight, blood pressure and temperature. The research interest is usually focussed on comparing mean difference or ratio between treatment and control groups (Sutton et al., 2000). The effect size measures for continuous data are grouped into two families, the d (difference) and r (relationship). We begin with the d family, starting with mean difference.

2.2.1.1 Mean difference

The mean difference measures the amount by which a treatment intervention changes the outcome on average compared with the control. It is useful when different studies outcomes are measured on the same scale.

Consider a study in which the outcomes are measured as means in two groups, treatment and control, and the focus is to compare the means. Let μ_t and μ_c be the means of the treatment and control groups estimated by the sample means $\bar{X}_t$ and $\bar{X}_c$, respectively. The mean difference is given by

$$\vartheta = \mu_t - \mu_c, \text{ estimated by } \hat{\vartheta} = \bar{X}_t - \bar{X}_c. \tag{2.3}$$

Its variance is given by

$$\text{var}(\vartheta) = \tfrac{\sigma_t^2}{n_t} + \tfrac{\sigma_c^2}{n_c}, \text{ estimated by } \text{var}(\hat{\vartheta}) = \tfrac{S_t^2}{n_t} + \tfrac{S_c^2}{n_c}, \tag{2.4}$$

where σ_t^2 and σ_c^2 are the variances of the treatment and control groups estimated by their sample variances S_t^2 and S_c^2, respectively, and n_t and n_c are the sample sizes of the treatment and control groups, respectively. The mean difference has the advantage of easy computation and it is easily interpretable.

2.2.1.2 Standardized mean difference

The mean difference effect size depends on the units of the measurements of the studies outcomes, and thus can not provide any meaningful information to the researcher when scale differs across the studies. To address this problem, the mean difference in (2.3) is scaled by dividing it with an appropriate standard deviation to obtain the standardized mean difference effect size measure. The standardized mean difference conveys the size of the effect relative to the variance in the sample data. The main assumption is that the variance is constant across the groups; $\sigma_c^2 = \sigma_t^2 = \sigma^2$. There are three different ways described below to define the standard deviation in the denominator.

Glass's Delta

According to Glass (1976), the most reasonable procedure to calculate the effect size is to divide the mean difference by the control group standard deviation. His argument was that pooling two variances could lead to different standardized values of the identical mean difference within an experiment where several treatments were compared to a control (Hedges, 1981). Glass's delta is given by

$$\Delta = \tfrac{\mu_t - \mu_c}{\sigma_c}, \text{ estimated by } \hat{\Delta} = \tfrac{\bar{X}_t - \bar{X}_c}{S_c}, \tag{2.5}$$

where S_c is the estimate of the control group standard deviation. The variance of Glass's Delta is calculated by

$$\mathrm{var}(\Delta) = \frac{n_t + n_c}{n_t n_c} + \frac{\Delta^2}{2(n_c - 2)}. \tag{2.6}$$

The differences in sample variances across groups can introduce bias in the effect size estimate thereby makes the Glass's Delta unreliable. We do not pursue this measure further.

Cohen's d

This effect size measure was proposed by Cohen (1988) as the mean difference divided by the pooled standard deviation to correct the likely bias in the Glass's Delta effect size estimate. The Cohen's d is given by

$$d = \tfrac{\mu_t - \mu_c}{\sigma}, \text{ estimated by } \hat{d} = \tfrac{\bar{X}_t - \bar{X}_c}{S_p}, \tag{2.7}$$

where

$$S_p = \sqrt{\frac{(n_t - 1)S_t^2 + (n_c - 1)S_c^2}{n_t + n_c}}. \tag{2.8}$$

The variance of the Cohen's d is given by

$$\text{var}(d) = \frac{n_t - n_c}{n_t n_c} + \frac{d^2}{2(n_t - n_c)}. \tag{2.9}$$

Hedges (1981) g

This is another alternative standardized mean difference estimator of the effect size given by

$$g = \frac{\bar{X}_t - \bar{X}_c}{S_{pH}}, \tag{2.10}$$

where S_{pH} is the pooled standard deviation suggested by Hedges (1981) as

$$S_{pH} = \sqrt{\frac{(n_t - 1)S_t^2 + (n_c - 1)S_c^2}{n_t + n_c - 2}}. \tag{2.11}$$

The Hedges's g is biased especially when the sample size is small. Hedges and Olkin (1985) proposed a correction factor given by

$$J = \left(1 - \frac{3}{4(n_t - n_c) - g}\right), \tag{2.12}$$

and the approximately unbiased estimator is given by

$$g^* = g \times J. \tag{2.13}$$

In general, the standardized mean difference effect size is easy to calculate and has consistent interpretation across different research studies. However, the unit-less values of its estimates require a more sophisticated acquaintance with the details of the application (Gibbons et al., 1977). Moreover, standardized mean difference effect sizes are only useful when research findings are not required to be expressed in the units of their measurement.

2.2.1.3 r family

The r or correlation family of effect sizes includes measures of the association between two variables. Correlation is well known to many researchers and is the most widely used effect size measure (Field, 2005), especially when research interest is in the relationship between variables in the treatment and control groups. The r family include the Pearson's product moment correlation (r) when both variables are continuous, the phi coefficient(ϕ) when both variables are dichotomous, point biserial coefficient (r_{pb}) when one variable is continuous and one is dichotomous, and the Spearman's rank correlation coefficient (rho (ρ)) when both variables are ranked.

When the population correlation is close to 1 the distribution of its sample estimates becomes skewed (Rosenthal et al., 1994) and this makes the correlation r unstable. Also, it makes the combination and interpretation of correlations difficult and complicated. The Fisher's transformation, $Z = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$ is usually used to stabilize the variance of the sample correlation. The distribution of Z is approximately normal with variance equal to $\sigma^2(Z) = (n-3)^{-1}$, where n is the sample size. The Fisher's transformation, Z is the effect size measure usually used in meta-analysis of Pearson correlations.

2.2.2 Effect size measures for binary data

A binary outcome is a response which assumes one out of two values. These values may be in the form of yes or no, agree or disagree, success or failure, effective or ineffective, exposed or unexposed, conform or not conform, etc. In many sciences, experiments are often conducted to compare treatment and control groups with the outcomes measured on binary scale. Table 2.1 is an example of a contingency table showing how binary outcomes from comparative studies can be summarised. In the next sections, the common effect size measures for binary data used in meta-analysis are presented.

2.2.2.1 Risk difference

The risk difference is an important effect size measure often used in meta-analysis. It describes the absolute changes in the risk that are attributed to the treatment arm. It is simply the difference between the probabilities of an event in two groups. Consider an

Table 2.1: An example of a contingency table for binary outcomes in comparative studies

	No. having event	No. not having event	Sample size	P(event)
Treatment Group	X	$n_t - X$	n_t	$\widehat{\mathrm{P}}_t = X/n_t$
Control Group	Y	$n_c - Y$	n_c	$\widehat{\mathrm{P}}_c = Y/n_c$

experiment in which two groups, treatment (t) and control (c) are compared in respect to outcomes measured on a binary scale. Let P_t and P_c denote the probabilities of the event, and n_t and n_c be the sample sizes of the groups, respectively. The risk difference effect size is estimated by

$$\widehat{\mathrm{RD}} = \widehat{P}_t - \widehat{P}_c, \tag{2.14}$$

and its variance is given by

$$\mathrm{var}(\mathrm{RD}) = \frac{\mathrm{P}_t(1-\mathrm{P}_c)}{n_t} + \frac{\mathrm{P}_c(1-\mathrm{P}_t)}{n_c}. \tag{2.15}$$

Risk difference is the simplest procedure for estimating the effect from binary outcomes. However the range of its variability is restricted by the magnitude of the probabilities P_t and P_c (Hedges et al., 1999) which is a major disadvantage to the effect size.

2.2.2.2 Relative risk

Relative risk is widely used in medicine because it is easy to understand and interpreted by both clinicians and the patients. It is simply the ratio between the probabilities of an events in the treatment and control groups. Relative risk is estimated by

$$\widehat{\mathrm{RR}} = \widehat{\mathrm{P}}_t/\widehat{\mathrm{P}}_c, \tag{2.16}$$

and it takes values from 0 to ∞. The variance of the log of its estimate is approximately

$$\mathrm{var}(\log \widehat{\mathrm{RR}}) \approx \frac{1-\mathrm{P}_t}{n_t P_t} + \frac{1-P_c}{n_c P_c}. \tag{2.17}$$

2.2.2.3 Odds ratio

The odds is another form of expressing probabilities, and it is widely used in gambling. The odds is the ratio of the probability that an event of interest occurs to the probability that it does not occur. When binary experimental outcomes are generated from two treatment arms, the effect size can be measured by the ratio of the odds of the event of interest between the two groups and the parameter is called the odds ratio. Odds ratio is estimated by

$$\widehat{\gamma} = \frac{\widehat{\mathrm{P}}_t/(1-\widehat{\mathrm{P}}_c)}{\widehat{\mathrm{P}}_c/(1-\widehat{P}_t)}, \tag{2.18}$$

and the variance of the sample log odds ratio is approximately

$$\mathrm{var}(\log \widehat{\gamma}) = \frac{1}{n_t \mathrm{P}_t} + \frac{1}{n_t(1-\mathrm{P}_t)} + \frac{1}{n_c \mathrm{P}_c} + \frac{1}{n_c(1-\mathrm{P}_c)}. \tag{2.19}$$

Remark 2.2.1. *It is important to note that we do not provide a detailed discussion of the effect size measures presented above. The usage of any of the effect size measure depends on the objectives of the study, the practical importance and the type of scale in which the studies outcomes were measured. Effect size measures should be chosen in such a way that the results of meta-analyses are easily interpretable and comparable across all studies.*

2.3 Models for combing results in meta-analysis

A fundamental issue in meta-analysis is the choice of an appropriate model that describes the underlying effect sizes from the different studies. There are two models used to combine results in meta-analysis; the fixed- and random-effects models (Hedges and Vevea, 1998, Hunter and Schmidt, 2000, Sutton et al., 2000). These models use different assumptions that lead to a different calculation and interpretation of the combined effect.

2.3.1 Fixed effect model

Fixed effect model (FEM) of meta-analysis assumes that all the included studies investigate the same population and therefore share a common location parameter. Denote by y_1, y_2, ..., y_K the estimates of treatment effects derived from K studies. When y_i's are sample means or mean difference, the fixed effect model is given by

$$y_i = \theta + e_i, \tag{2.20}$$

where θ is the common location parameter, $e_i \sim N(0, \sigma_i^2)$ is the sampling error, σ_i^2 are the within-study variances, for i=1, 2, ..., K. For other effects measures, approximate normality of y_i's holds when the sample sizes n_i of the studies are relatively large. Appropriate estimates S_i^2 of the variances σ_i^2 are easily calculated for all effect sizes used in meta-analysis and are habitually treated as known constants (Viechtbauer, 2007). In FEM, each study is assigned a weight proportional to the inverse of the within-study variance, which is denoted by $w_i = 1/S_i^2$. The combined effect is estimated as a weighted mean of the individual effect estimates given by

$$\hat{\theta}_{FEM} = \sum_i w_i y_i / W, \tag{2.21}$$

where $W = \sum_{i=1}^{K} w_i$. The variance of the combined effect is given by the inverse of the sum of weights, W^{-1}.

2.3.1.1 Inference in FEM

Standard inference in FEM is based on approximate normality of the distribution of the combined effect, $\hat{\theta}_{FEM} \sim N(\theta, W^{-1})$. Therefore the confidence intervals of the population treatment effect are given by

$$\hat{\theta}_{FEM} \pm z_{1-\alpha/2} W^{-\frac{1}{2}}. \tag{2.22}$$

To test the hypothesis for the presence or otherwise of a treatment effect, $H_0 : \theta = 0$ against $H_1 : \theta_1 \neq 0$, the Wald's statistic

$$Z_W = W^{\frac{1}{2}} |\hat{\theta}_{FEM}| \tag{2.23}$$

is compared with the critical values for the standard normal distribution.

Often, in order to test the hypothesis of homogeneity of treatment effects, $H_0 : \theta_1 = \theta_2 = ... = \theta_K = \theta$ against $\theta_i \neq \theta_j$, for some $i \neq j$, the Cochran Q statistic

$$Q = \sum_{i=1}^{K} w_i (y_i - \hat{\theta}_{FEM})^2. \tag{2.24}$$

plays an important role in meta-analysis. It is widely used in inference on heterogeneity of treatment effects. The Q statistic is routinely assumed to follow the chi-square distribution with $K-1$ degrees of freedom χ^2_{K-1}, though this is true only for very large sample sizes, see Hoaglin (2015).

2.3.2 Random-effects model

Random-effects model (REM) is generally preferred to the fixed effect model (Hunter and Schmidt, 2000) due to its ability to account for variation across the studies. The random effects model allows generalisation of mean effects θ_i across studies and it assumes that they are sampled from a population of parameters with mean θ. Random-effects model is a two level model given by

$$\begin{aligned} y_i &= \theta_i + e_i;\ e_i \sim F(0, \sigma_i^2) \\ \theta_i &= \theta + \epsilon_i;\ \epsilon_i \sim G(0, \tau^2), \end{aligned} \tag{2.25}$$

where F and G come from an arbitrary short-scale families of distribution and σ_i^2 and τ^2 are the within- and between-study variances, respectively. The most popular choice is two normal distributions. Then marginally the random effects model is defined by

$$y_i = \theta + \xi_i;\ \xi_i \sim N(0, \tau^2 + \sigma_i^2). \tag{2.26}$$

The between-study variance, τ^2 describes the degree of inconsistency among the effect estimates. The special case where $\tau^2 = 0$ implies that the effect sizes, $\theta_1 = \theta_2 = ... = \theta_K$ are homogeneous (Viechtbauer, 2007), and the resulting model reduces to FEM in (2.20). The weights assigned to studies in REM are inverse variance weights defined by $w_i^* = w_i(\tau^2) = (\tau^2 + \sigma_i^2)^{-1}$. Estimated values of τ^2 and σ_i^2 are substituted in practice. Similar to FEM, the combined effect in REM is estimated as a weighted mean of the individual effect estimates, $\hat{\theta}_{REM} = \sum\limits_i w_i^* y_i / \sum\limits_i w_i^*$.

2.3.3 Mixed-effects meta-regression model

Effect estimates on some study-level covariates such as patients mix, time, climate change, etc. This increases heterogeneity between the studies effects, and it is difficult to use the standard random-effects model to describe the results. Meta-regression model allows results from studies to relate to study-level covariates. The mixed-effects meta-regression model is based on the assumption that $y_i|x_i \sim N(x_i\beta, \tau^2 + \sigma_i^2)$, where y_i is the estimated effects from the ith study, i =1, 2, ..., K, x_i is the vector of the study level covariates and β is a $p \times 1$ vector of regression parameters. Thus the model is described by

$$y_i = x_i\beta + \theta_i + e_i; \theta_i \sim N(0, \tau^2) \text{ and } ei \sim N(0, \sigma_i^2). \tag{2.27}$$

Equivalently, matrix form (Jackson et al., 2014) of this model is

$$Y|X \sim N(X\beta, \Delta + \tau^2 I), \tag{2.28}$$

where Y is a column vector containing the y_i, X is the $K \times p$ design matrix whose ith row is x_i, Δ is a diagonal matrix containing the σ_i^2 and I is $K \times K$ identity matrix.

2.3.3.1 Estimation of heterogeneity in treatment effects, τ^2

The between-study variance, τ^2 has a crucial role in assessing the degree of consistency of treatment effects across the studies (Higgins et al., 2003), and thus its estimation is an important issue in meta-analysis. This section introduces some of the common methods for estimating τ^2. Each method differs in terms of precision and bias in estimating τ^2, and therefore may have a different effect on sequential testing of the treatment effects. In Sections 4.3 and 5.2, the DerSimonian and Laird (1986), Higgins et al. (2011), Paule and Mandel (1982) and REML estimators of τ^2 are used to examine by simulation how this affects the sequential testing for random-effect meta-analysis.

Method of moments

Suppose, a_i are constants corresponding to the effect estimates y_i for i=1, 2, ..., and that the combined effect, $\hat{\theta} = \sum\limits_i a_i y_i / \sum\limits_i a_i$. Then the expected value is

$$\mathrm{E}\left\{\sum_i a_i(y_i - \hat{\theta})^2\right\} = \sum_i a_i(\tau^2 + \sigma_i^2/n_i) - \sum_i a_i^2(\tau^2 + \sigma_i^2/n_i)/\sum_i a_i, \tag{2.29}$$

where σ_i^2 and n_i are the variances and sample sizes, respectively (DerSimonian and Kacker, 2007). Equation (2.29) can be simplified to obtain

$$\mathrm{E}\left\{\sum_i a_i(y_i - \hat{\theta})^2\right\} = \tau^2\left\{\sum_i a_i - \sum_i a_i^2/\sum_i a_i\right\} + \left\{\sum_i a_i^2\sigma_i^2/n_i - \sum_i (a_i\sigma_i^2/n_i)/\sum_i a_i\right\}. \tag{2.30}$$

Substituting S_i^2 for σ_i^2 and solving (2.30) for τ^2, the moment estimator of τ^2 is given by

$$\hat{\tau}^2_{MM} = \frac{\sum\limits_i a_i(y_i - \hat{\theta})^2 - \left\{\sum\limits_i a_i S_i^2 - \sum\limits_i a_i^2 S_i^2/\sum\limits_i a_i\right\}}{\sum\limits_i a_i - \sum\limits_i a_i^2/\sum\limits_i a_i}. \tag{2.31}$$

The value of $\hat{\tau}^2_{MM}$ is constrained to non-negative values, and the constants a_i are usually chosen as the weights assigned to the studies.

The Cochran (1954) method

The Cochran (1954) method is a moment method where a fixed constant $1/K$ is assigned to each study as the weight, and K is the number of studies in the meta-analysis. The combined effect is determined by the arithmetic mean of the effect estimates, $\hat{\theta}_C = \sum_{i=1}^{K} y_i/K$. Substituting $\hat{\theta}_C$ for $\hat{\theta}$ and $1/K$ for a_i in (2.31), the Cochran (1954) estimator is defined by

$$\hat{\tau}_C^2 = \frac{1}{K-1}\sum_{i=1}^{K}(y_i - \hat{\theta}_C)^2 - \frac{1}{K}\sum_{i=1}^{K} S_i^2, \tag{2.32}$$

and is also constrained to non-negative values.

DerSimonian and Laird (1986) Method

The DerSimonian and Laird (1986) estimator can be calculated by substituting $w_i = n_i/\hat{\sigma}_i^2$ in (2.31). The estimator is given by

$$\hat{\tau}_{DL}^2 = \frac{Q-(K-1)}{C}, \tag{2.33}$$

where $Q = \sum_{i=1}^{K} w_i(y_i - \hat{\theta})^2$ and $C = \sum_{i=1}^{K} w_i - \frac{\sum_{i=1}^{K} w_i^2}{\sum_{i=1}^{K} w_i}$.

Higgins et al. (2011) Method

Higgins et al. (2011) proposed an estimator of τ^2 specifically for sequential testing. The estimator is modified from DerSimonian and Laird (1986) method using semi-Bayes approach and is defined by

$$\hat{\tau}_H^2 = \frac{2\lambda + K\hat{\tau}_{DL}^2}{2\eta + K - 2}, \tag{2.34}$$

where λ and η are parameters of a prior inverse gamma distribution for τ^2.

Paule and Mandel (1982) Method

Denote $w_i^*(\tau^2) = (\tau^2 + \sigma_i^2)^{-1}$, the weights assigned to studies in REM as a function of τ^2. Define $Q(\tau^2) = \sum w_i^*(\tau^2)(y_i - \hat{\theta}(\tau^2))^2$. The Paule and Mandel (1982) estimator of τ^2 is calculated from the solution of the estimating equation for the expected value of the Q statistic under H_0 given by

$$Q^2(\tau^2) - (K - 1) = 0, \tag{2.35}$$

The Paule and Mandel (1982) estimator is statistically optimal, in the sense that the estimator is not biased and has minimum variance, when the distribution of the effect estimates is normal. However the method does not generally require any normality assumptions (DerSimonian and Kacker, 2007).

Maximum likelihood method

All the methods discussed above are moment estimators with the exception of Paule and Mandel (1982) which is iterative. There are other alternative approaches for estimating τ^2 based on maximum and restricted maximum likelihood. The standard assumption in random-effects model is that the distribution of the effect estimates is normal, $y_i \sim N(\theta, \tau^2 + \sigma_i^2)$. The log-likelihood function of θ and τ^2 is then given by

$$l(\theta, \tau^2) = -\frac{1}{2}\sum \log(\tau^2 + \sigma_i^2) - \frac{1}{2}\sum \frac{(y_i - \theta)^2}{\tau^2 + \sigma_i^2} + C, \tag{2.36}$$

where C is a constant. Setting the partial derivatives with respect to θ and τ^2 equal to zero and solving the resulting equation, the maximum likelihood (ML) estimates of θ and τ^2 are given by

$$\hat{\theta}_{REM} = \sum_i w_i^* y_i / \sum_i w_i^* \text{ and } \hat{\tau}^2_{ML} = \frac{\sum w_i^{*2}[(y_i - \hat{\theta}_{REM})^2 - \sigma_i^2]}{\sum w_i^{*2}}, \tag{2.37}$$

where $w_i^* = (\tau^2 + \sigma_i^2)^{-1}$ is the weight assigned to studies in REM. The solution of (2.37) is determined iteratively starting with an initial value $\hat{\tau}^2_{REM} = \tau^2_0$, and should the result converge to a negative value, it is truncated at zero (Viechtbauer, 2007).

Restricted maximum likelihood method

In a finite sample the maximum likelihood estimator, $\hat{\tau}^2_{ML}$ underestimates the population heterogeneity (Corbeil and Searle, 1976), and it is negatively biased (Corbeil and Searle, 1976, Viechtbauer, 2005). The restricted maximum likelihood (REML) is the alternative approach to correct the underestimation. Its log-likelihood function is given by

$$l_R(\tau^2) = -\frac{1}{2}\sum \log(\tau^2 + \sigma_i^2) - \log \sum \frac{1}{\tau^2 + \sigma_i^2} - \frac{1}{2}\sum \frac{(y_i - \hat{\theta}_{REM})^2}{\tau^2 + \sigma_i^2} + C, \tag{2.38}$$

where C is a constant. Setting the partial derivative equal to zero and solving the resulting equation, the restricted maximum likelihood estimate is given by

$$\hat{\tau}^2_R = \frac{\sum w_i^{*2}[(y_i - \hat{\theta}_{REM})^2 - \sigma_i^2]}{\sum w_i^{*2}} + \frac{1}{\sum w_i^*}, \tag{2.39}$$

and also computed iteratively in the same manner as in $\hat{\tau}^2_{ML}$.

2.3.3.2 Confidence intervals for τ^2

The confidence interval (CI) for τ^2 indicate the precision with which the heterogeneity variance is estimated. It contains important information for the associated analysis of heterogeneity (Viechtbauer, 2007). Several methods for constructing the confidence interval for τ^2 have been proposed, but presented here are only a few that can be applied to the estimators of τ^2 discussed in Section 2.3.3.1.

Wald-type confidence intervals for τ^2

From the maximum and restricted maximum likelihood functions of τ^2 in equations (2.36) and (2.38), respectively, it can be shown (Viechtbauer, 2007, Rao et al., 1981) that the variances, $\text{var}(\hat{\tau}^2_{ML}) = 2\sum w_i^2$ and $\text{var}(\hat{\tau}^2_{REML}) = 2\left(\sum w_i^2 - 2\frac{\sum w_i^3}{\sum w_i} + \frac{(\sum w_i^2)^2}{(\sum w_i)^2}\right)^{-1}$. Based on the asymptotic normality of $\hat{\tau}^2_{ML}$ and $\hat{\tau}^2_{REML}$ the $100(1-\alpha)\%$ Wald confidence intervals for τ^2 are given (Biggerstaff and Tweedie, 1997, Viechtbauer, 2007) by

$$\hat{\tau}^2_{ML} \pm z_{1-\alpha/2}\sqrt{2\sum w_i^2} \tag{2.40}$$

and

$$\hat{\tau}^2_{REML} \pm z_{1-\alpha/2}\sqrt{2\left(\sum w_i^2 - 2\frac{\sum w_i^3}{\sum w_i} + \frac{(\sum w_i^2)^2}{(\sum w_i)^2}\right)^{-1}}, \tag{2.41}$$

where $z_{(1-\alpha)}$ is the $100(1-\alpha/2)$th percentile of the standard normal distribution.

Biggerstaff and Tweedie (1997) method

The Biggerstaff and Tweedie (1997) method is based on the Cochran Q-statistic given in equation (2.24) and used for constructing CI for τ^2 estimated by DerSimonian and

Laird (1986) method. The expected value and variance of the Q-statistic are given by

$$\mathrm{E}[Q] = (K-1) + \left(S_1 + \tfrac{S_2^2}{S_1}\right)\tau^2 \text{ and } \mathrm{var}[Q] = 2(K-1) + 4\left(S_1 + \tfrac{S_2^2}{S_1}\right)\tau^2 + 2\left(S_2 + 2\tfrac{S_3}{S_1} + \tfrac{S_2^2}{S_1^2}\right)\tau^4, \tag{2.42}$$

where $S_j = \sum (w_i)^j$ and K is the number of studies. Using this, Biggerstaff and Tweedie (1997) proposed to approximate distribution for τ^2 by a gamma distribution with shape and scale parameters given by $\gamma(\tau^2) = \frac{(\mathrm{E}[Q])^2}{\mathrm{var}[Q]}$ and $\phi(\tau^2) = \frac{\mathrm{var}[Q]}{\mathrm{E}[Q]}$, respectively. Let $f(y/\gamma(\hat{\tau}^2))$ be the approximate density of $\hat{\tau}^2$, the $1-\alpha$ percent CI is the obtained (Biggerstaff and Tweedie, 1997, Viechtbauer, 2007) by finding the two values of $\hat{\tau}^2$ that satisfy the following equation

$$\int\limits_{Q/\phi(\tau^2)}^{\infty} f(y/\gamma(\hat{\tau}^2)) = \alpha/2. \tag{2.43}$$

Profile likelihood method

The profile likelihood method uses the contour plots of the profile likelihood of τ^2 to construct CI, see Viechtbauer (2007), Hardy and Thompson (1996). A contour plot is a two dimensional plot that shows one-dimensional curves, called contour lines. In other words it is a plot that displays 3-dimensional relationship in two dimensions. For example, a 95% CI of τ^2 can be obtained when contour plots of the profile likelihood of τ^2 satisfy the equation

$$L(\tau^2) > L(\hat{\tau}^2) - 3.84/2, \text{ for } L(\tau^2) = -\tfrac{1}{2}\sum \ln(\tau^2+\sigma_i^2) - \tfrac{1}{2}\ln\sum\left(\tfrac{1}{\tau^2+\sigma_i^2}\right) - \tfrac{1}{2}\sum\tfrac{(y_i-\hat{\theta})^2}{\tau^2+\sigma_i^2}, \tag{2.44}$$

where 3.84 is the 5% point of the χ_1^2 distribution (Viechtbauer, 2007, Hardy and Thompson, 1996).

Bootstrap confidence intervals

Viechtbauer (2007) proposed the use of parametric and non-parametric bootstrap procedures for obtaining the CIs of τ^2. In this method, a set of B bootstrap estimates $\{\hat{\tau}_b^2$: b=1, 2, ..., $B\}$ of τ^2 are obtained from B bootstrap samples of the data. Then ordering the set $\{\hat{\tau}_b^2$: b=1, 2, ..., $B\}$ in ascending order the $1-\alpha$ percent CI's are given by the $(100\alpha/2)$th and $100(1-\alpha/2)$th empirical percentiles of the $\hat{\tau}_b^2$.

2.3.3.3 Inference in random-effects model

As in FEM, standard inference in random-effects model is based on the asymptotic normality of the combined effect, $\hat{\theta}_{REM} \sim N\left(\theta, (\sum w_i^*)^{-1}\right)$. The confidence intervals are defined by

$$\hat{\theta}_{REM} \pm z_{1-\alpha/2}\left(\sum w_i^*\right)^{-\frac{1}{2}} \tag{2.45}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)^{th}$ percentile of the normal distribution. Due to the addition of the heterogeneity variance in REM, its confidence intervals are wider in comparison to FEM. Therefore the inference in REM is more conservative in terms of statistical significance of the combined effect.

To test for the presence or otherwise of a treatment effect, the Wald statistic,

$$Z_W = \left(\sum w_i^*\right)^{\frac{1}{2}} |\hat{\theta}_{REM}| \tag{2.46}$$

is compared with the critical values of the standard normal distribution. The Q statistic given in (2.24) is used to test the hypothesis of heterogeneity for the existence of the variance component τ^2, $H_0 : \tau^2 = 0$ vs $H_1 : \tau^2 > 0$.

Chapter 3

Review of sequential methods and methods for monitoring trends in meta-analysis

In recent years, temporal changes in effect size have been reported in many fields of research. Examples are given in Hodgson et al. (1989); Nieuwkamp et al. (2009); Hyde et al. (1990); Twenge. et al. (2008); Gehr et al. (2006); Grabe et al. (2008). Temporal changes in effect sizes present a serious danger to the validity of results and conclusions of meta-analysis, and thus several methods have been proposed to monitor the trends so that results and conclusions in meta-analysis can be interpreted based on the time it was conducted. This Chapter reviews the common methods used for monitoring the temporal trends in magnitude of effect sizes in meta-analysis with special focus on sequential methods in meta-analysis. The first Section introduces sequential analysis.

The second Section reviews the standard sequential methods in meta-analysis including 'simplistic methods' for monitoring trends in meta-analysis. The third Section focuses on adaptive methods for sequential decisions in clinical trials.

3.1 Sequential analysis

Sequential analysis was initially introduced during World War II in response to the overwhelming demands for methods of testing the efficiency of aircraft gunnery (Lai, 2001). In this method of analysis the sample size is not fixed in advance, instead data are evaluated as more observations are collected, and further sampling is stopped in accordance with a predefined stopping rule as soon as significant result is observed. Sequential methods are popular in many areas where sequential monitoring of process outputs is required. For example, they are used in engineering, monitoring of prices of goods and services and quality control. In meta-analysis, sequential methods are increasingly becoming popular, see Whitehead (1997a); Higgins et al. (2011); Bollen et al. (2006); van der Tweel and Bollen (2010) as examples. They are used for gauging sufficiency in accumulating evidence (Lau et al., 1992, Pogue and Yusuf, 1997, Wetterslev et al., 2008) and serve as an appropriate statistical tool to monitor any possible trends in meta-analysis. This Section introduces three sequential methods; the sequential probability ratio test, the CUSUM scheme and group sequential methods. The properties and the advantages of the sequential methods in the context of meta-analysis are also highlighted.

3.1.1 Sequential probability ratio test (SPRT)

Introduced by Wald (1945), SPRT is a sequential test for a simple null hypothesis H_0 against a simple alternative hypothesis H_1. It is based on a likelihood ratio which is treated as a function of the observations. Consider a sequence of independent and identically distributed random variables X_1, X_2, X_3, ... with the same probability density function, $f(X)$. To test the null hypothesis $H_0 : f = f_0$ against the alternative hypothesis $H_1 : f = f_1$, the SPRT stops sampling at the stage

$$N = \inf\left\{ n > 1 : \lambda_n \notin (A, B) \right\}, \tag{3.1}$$

where

$$\lambda_n = \prod_{i=1}^{n} \frac{f_1(X_i)}{f_0(X_i)} \tag{3.2}$$

is the likelihood ratio at stage n and $0 < A < B < \infty$ are stopping boundaries. When stopping occurs, decisions are taken as follows. If $\lambda_n \leq A$ decide H_0, if $\lambda_n \geq B$ decide H_1. The choice of the stopping boundaries, A and B depends on the pre-specified Type I and II error probabilities.

3.1.1.1 Stopping boundaries and the error probabilities

The decision not to reject or reject a statistical hypothesis depends on the costs associated with committing an error (Hubbard and Bayarri, 2003). This could be a Type I or Type II error. The Type I error (false rejection) is the probability of deciding H_1 when H_0 is true; while the Type II error (false acceptance) is the probability of deciding H_0 when H_1 is true. To establish their relationship with the stopping boundaries, denote the Type I and Type II error probabilities by $\alpha = \mathrm{P}(H_1|H_0)$, the probability of H_1 given that H_0 is true and $\beta = \mathrm{P}(H_0|H_1)$, the probability of H_0 given that H_1 is true, respectively. Let $X = (x_1, x_2, ..., x_n)$ and $p_j(X) = \prod_{i=1}^{n} f_j(x_i)$, j=(0,1). Define the decision sets R_0=$\{(x_1, ..., x_n); N = n$ and $\lambda_N \leq A\}$ and R_1=$\{(x_1, ..., x_n); N = n$ and $\lambda_N \geq B\}$. The power of the test is given by

$$\begin{aligned} 1 - \beta =& \mathrm{P}(H_1|H_1) \\ =& \int\limits_{R_1} p_1(X)dX \\ =& \int\limits_{R_1} \frac{p_1(X)}{p_0(X)} p_0(X)dX \\ =& \int\limits_{R_1} \lambda_n p_0(X)dX \\ \geq& B \int\limits_{R_1} p_0(X)dX \\ =& B\mathrm{P}(H_1|H_0) \\ =& B\alpha. \end{aligned} \tag{3.3}$$

Similarly,

$$\begin{aligned} 1-\alpha =&1-\mathrm{P}(H_1|H_0) \\ =&\mathrm{P}(H_0|H_0) \\ =&\int\limits_{R_0} p_0(X)dX \\ =&\int\limits_{R_0} \lambda_n^{-1}p_1(X)dX \\ \geq& A^{-1}\int\limits_{R_0} p_1(X)dX \\ =&A^{-1}\mathrm{P}(H_0|H_1) \\ =&A^{-1}\beta \end{aligned} \tag{3.4}$$

Treating the inequalities of (3.3) and (3.4) as approximate equalities and solving for A and B, the stopping boundaries are defined by

$$A = \beta/(1-\alpha) \text{ and } B = (1-\beta)/\alpha. \tag{3.5}$$

See the detailed derivation in Lauritzen (2004) and Nowak (2011).

3.1.1.2 Sample size of SPRT

The expected sample size of SPRT is the average number of observations required before a decision is arrived at. The formula of the expected sample size for the SPRT is derived based on the Wald's equation, see Nowak (2011). The Wald's equation states that if N is a stopping time with respect to an independent and identically distributed sequence $\{X_n : n \geq 1\}$, and if $\mathrm{E}[N] < \infty$ and $\mathrm{E}[|X|] < \infty$ then

$$\mathrm{E}\left\{\sum_{n=1}^{N} X_n\right\} = \mathrm{E}[N]\mathrm{E}[X] \tag{3.6}$$

Since X_1, X_2, ... are independent and identically distributed random variables, the logarithm of the likelihood statistic in (3.2)

$$\log(\lambda_n) = \sum_{i=1}^{n} \log(f_1(X_i)/f_0(X_i)) \tag{3.7}$$

is a sum of independent and identically distributed variables. Let N be the first $n \geq 1$ such that $\log(\lambda_n) \notin (a, b)$, where $a = \log(A)$ and $b = \log(B)$. Therefore, by (3.6),

$$\mathrm{E}\left[\log(\lambda_N)\right] = \mu_j \mathrm{E}_j[N], \tag{3.8}$$

where $\mu_j = \mathrm{E}_j\left[\log f_1/f_0 \middle| H_j\right]$, for j=0, 1. But $\mathrm{E}[\log(\lambda_N)] \approx a\Pr(\lambda_n \leq A) + b\Pr(\lambda_n \geq B)$, see Siegmund (1985). It follows that the expected sample size of the SPRT under the null and alternative hypotheses are given by

$$\mathrm{E}_0[N] = \mu_0^{-1}\left\{\alpha\log\left(\frac{1-\beta}{\alpha}\right) + (1-\alpha)\log\left(\frac{\beta}{1-\alpha}\right)\right\} \tag{3.9}$$

and

$$\mathrm{E}_1(N) = \mu_1^{-1}\left\{(1-\beta)\log\left(\frac{1-\beta}{\alpha}\right) + \beta\log\left(\frac{\beta}{1-\alpha}\right)\right\}, \tag{3.10}$$

where $\mu_0 = \mathrm{E}_0[\log(f_1/f_0)]$ and $\mu_1 = \mathrm{E}_1[\log(f_1/f_0)]$, (see Siegmund (1985)).

For the case of testing a simple null hypothesis against a simple alternative hypothesis, Wald and Wolfowitz (1948) showed that the SPRT leads to optimal solution of testing H_0 against H_1, in the sense that it minimizes both $\mathrm{E}_0[N]$ and $\mathrm{E}_1[N]$ among all tests whose sample size N has a finite expectation under H_0 and H_1 with error probabilities satisfying

$$\mathrm{P}_0(\text{Reject } H_0) \leq \alpha \text{ and } \mathrm{P}_1(\text{Reject } H_1) \leq \beta. \tag{3.11}$$

Remark 3.1.1. *The optimality and the applicability of SPRT to any observations with known distribution are desirable properties in sequential testing. However the SPRT is designed for testing a simple null hypothesis against simple alternative, and therefore its optimality property is restricted to such situations (Lai, 2001, Siegmund, 1985). This lack of optimality in general may have serious consequences that can prevent its uses in meta-analysis.*

3.1.2 The cumulative sum (CUSUM) scheme

Shewhart (1931) introduced statistical quality control charts, a class of sequential methods for monitoring and evaluation of the quality of products from continuous production line. In his monitoring scheme, a statistic is computed from fixed size samples of observations taken at regular intervals, which is then compared to predefined monitoring boundaries. If the value of the statistic is within the boundaries, the process is said to be in control. If the current value crosses the boundaries, then the process is out of control and corrective measures need to be taken to put the process back under control. However Shewhart (1931) sequential monitoring chart is a “single sample” scheme with the decision solely depending on the current sample although the results of previous samples are available on the chart (Lai, 2001). Motivated by these shortcomings, Page

(1954) modified the SPRT to develop the CUSUM scheme. The CUSUM scheme is a good alternative procedure for monitoring process when detection of small changes is important. It has the advantage of incorporating all information in the sequence of values (Montgomery, 2000).

Consider a sequence of independent and identically distributed observations, X_1, X_2, X_3, ... and a reference point k chosen between the target value μ_0 and the value corresponding to a point in the observations considered to be just unsatisfactory μ_1. The CUSUM statistic is computed by summing the successive differences $X_i - k$ for i=1, 2, 3, ... to build up a series given by

$$S_n = \sum_{i=1}^{n} (X_i - k). \tag{3.12}$$

A graph of the statistic S_n against the sample number (time) is called the CUSUM chart. If the path of the CUSUM chart is moving horizontally, then the process is in control. But if at any point of sampling the path rises significantly above or falls below the target value, μ_0 then the process is said to be out of control and something must be done to correct the process. A CUSUM chart designed to detect an upward trend is called upper CUSUM and the CUSUM chart for detecting downward trend is called lower CUSUM. The recursive formulae starting from zero for computing the upper and the lower CUSUMs are given by

$$S_n^+ = \max\left\{0, X_n - (\mu_0 + k) + S_{n-1}^+\right\} \text{ and } S_n^- = \max\left\{0, (\mu_0 - k) - X_n + S_{n-1}^-\right\}, \tag{3.13}$$

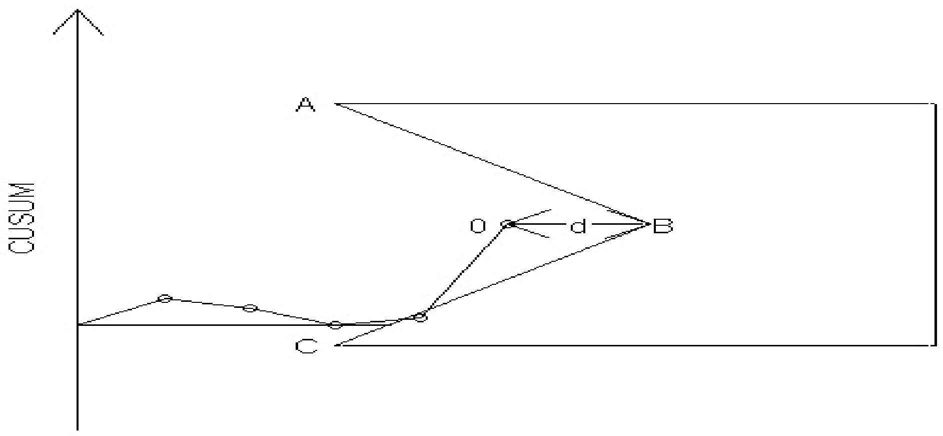


Figure 3.1: V-mask Procedure

respectively. The constant $k = (\mu_1 - \mu_0)/2$ is a reference point chosen mid-way between the mean target value μ_0 and the value of the parameter μ_1 considered to be just unsatisfactory.

3.1.2.1 Methods for detecting a change in mean of a process using a CUSUM scheme

The common methods of deciding when there is a change in monitored process using the CUSUM scheme include decision interval scheme, V-mask procedure and the CUSUM procedure.

V-mask procedure

A visual method for deciding when a change in monitored process occurred using the CUSUM is the V-mask procedure described by Barnard (1959). This method superimposes a V-shaped mask on the CUSUM chart in such a way that the vertex of the V-mask is pointed forward at a distance d (lead distance) from the latest point on the chart. The angle between the two arms of the V-mask is 2θ, where $\theta = \angle ABO = \angle OBC$ is the angle between an arm of the V-mask and the horizontal axis, see Figure 3.1. Performance of the V-mask procedure is measured by the lead distance, d and the angle θ, which are often chosen empirically (Wieringa et al., 1999). As long as the plotted values of the CUSUM remain within the arms of the V-mask, the process is in control. However if a point on the chart reaches or crosses the arms, then the process will be considered to have gone out of control. The probability of Type I error in the V-mask procedure is proportional to the lead distance, d and the angle θ (Montgomery, 2000, Woodward and Goldsmith, 1967). See Barnard (1959) and Wieringa et al. (1999) for more discussion of this method.

Decision interval scheme

In the decision interval scheme, a reference point k is chosen mid-way between the value of the parameter under the null hypothesis, μ_0 and the value of the parameter under the alternative hypothesis, μ_1. Then say for an upper CUSUM, as long as the S_n^+ is less than k, the process is in control. If at any point in sampling the value of S_n^+ reaches or crosses the reference point k, a CUSUM chart is started. When the CUSUM reaches or crosses the decision interval line h, then it is concluded that the process has gone out of control. The general detection criterion for this scheme is given by

$$S_n^+ \geq h \text{ or } S_n^- \leq -h \tag{3.14}$$

A reasonable choice for the value of the decision interval line h is usually set at $h = 5\sigma$, where σ is standard deviation of the observation, see Montgomery (2000). However, the decision interval line can also be determined to satisfy a desired average run length ARL (which we shall discuss later) using a method called the thumb rule. For detecting a shift with magnitude $\mathit{\Delta} = \mu_1 - \mu_0 \neq 0$ and $\mathit{\Delta} > k$, the thumb rule is defined by

$$ARL(\mathit{\Delta}) = 1 + \frac{h}{\mathit{\Delta} + k}. \tag{3.15}$$

Shu and Jiang (2006) derived a relationship between the ARL approximation by Siegmund (1985) and the the decision interval line given by

$$h = \frac{\log\{1 + 2k^2 ARL_0 + 2.332k\}}{2} - 1.166. \tag{3.16}$$

Figure 3.2 is an example of a two-sided decision interval scheme. See Montgomery (2000), Wieringa et al. (1999), Woodward and Goldsmith (1967) for a detailed description of the decision interval scheme.

CUSUM procedure

The Page (1954) CUSUM procedure is a special case of repeated sequential probability ratio test (RSPRT), where the stopping boundaries $a = \log A = 0$ and $b = \log B = h$.

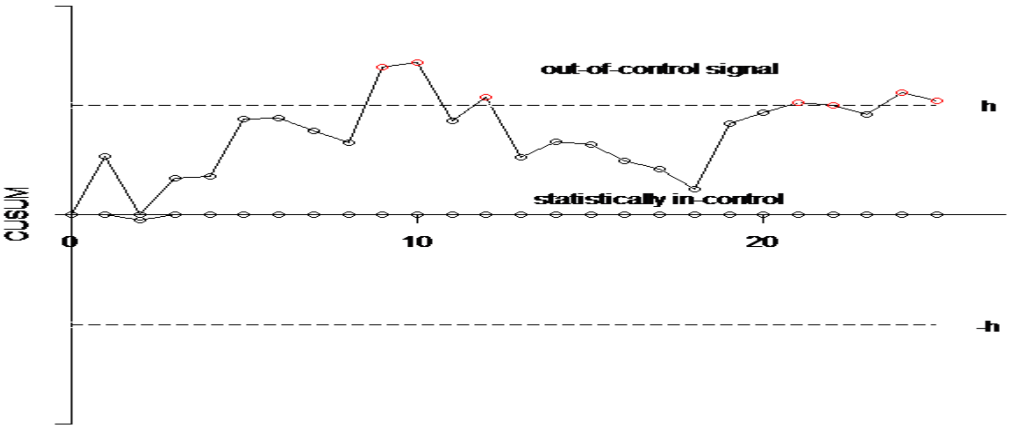


Figure 3.2: Decision Interval Scheme with the red points above the upper interval line (h) indicate out-of-control signal in the upper CUSUM.

In SPRT the decision rule is a function of the exit time, $N = \inf\{n > 1 : \lambda_n \notin (A, B)\}$ and $\lambda_n = \prod_{i=1}^{n} \frac{f_1(X_i)}{f_0(X_i)}$, see (3.1) and (3.2). At the exit time when $\lambda_n \geq A$, the SPRT terminates in favour of H_0. But in the CUSUM procedure instead of terminating the test, the SPRT is continually restarted as long as the decision favours H_0 until the time when $\lambda_n \geq B$ and the decision favours H_1. This is to say that the CUSUM procedure is a repeated SPRT. To define the CUSUM procedure, consider the following hypotheses.

$$\begin{aligned} &H_0 : X_1, ..., X_n \sim f_0 \\ &H_1 : X_1, ..., X_{K-1} \sim f_0 \text{ and } X_K, ..., X_n \sim f_1, \end{aligned} \tag{3.17}$$

where K is an unknown time of change. The likelihood ratio for the hypotheses is given by

$$\lambda_{n,K} = \frac{\prod_{i=1}^{K-1} f_0(X_i) \prod_{i=K}^{n} f_1(X_i)}{\prod_{i=1}^{n} f_0(X_i)}, \tag{3.18}$$

and the maximum likelihood ratio is

$$\begin{aligned} S_{nK} =& \sum_{i=K}^{n} \log\left(f_1(X_i)/f_0(X_i)\right) \\ =& S_n - ^{\min}_{0\le K\le n} S_K. \end{aligned} \tag{3.19}$$

Page (1954) proposed the stopping rule for the test as

$$N = \left\{n : S_n - ^{\min}_{0\le K\le n} S_K \ge h\right\}, \tag{3.20}$$

where h is determined by the stopping boundaries of SPRT such that $a = \log A = 0$ and $b = \log B = h$.

The optimality of the CUSUM in terms of optimal stopping time have been well established. See for example Moustakides (1986), who proved its optimality property similarly to the optimality in sequential probability ratio test.

3.1.2.2 The average run length (ARL) of a CUSUM

The average run length (ARL) is used as a major criterion for selecting a suitable CUSUM procedure (Woodall, 1983) as well as a tool for evaluation of its performance. It is the average number of observations required before a CUSUM scheme signals an alarm for a change. A high ARL should be expected when the process is operating at a satisfactory level and low when it is operating at unsatisfactory level. There are numerous ways to calculate the ARL, however we present here the Page (1954) integral method and Siegmund (1985) approximation of the ARL.

Page (1954) integral method:

In equation (3.12), H_0 is accepted when $S_n = \sum_{i=1}^{n}(X_i - k)$ is less than or equal to zero, and therefore only the first X such that $(X - k) > 0$ is necessary to accumulate scores. Let z be the first score such that it is bounded by 0 and h. The CUSUM is a repeated SPRT and terminates at the first time when a test crosses the boundary line h. In this case a single SPRT is defined by a path starting at z and ending either at 0 or at h. For a new observation X, the current score results to $z + X - k$, provided it belongs to the open interval $(0, h)$. If $z + X - k \leq 0$, the test stops and a new test is started from 0, and for $z + X - k \geq h$ the test stops and decision is taken. Let $\mathrm{P}(z)$ denote the probability that a test starts at z and ends at point $z \leq 0$. Denote $N(z)$ as the average sample number of the test starting at z. Let $L(z)$ denote the ARL of the CUSUM that starts at z, but all subsequent tests start at 0. Let $f(X)$ denote the probability density of the observations, and $F(x)$ be the cumulative distribution. For a single test that starts the score z the probability of the first event is $\mathrm{P}(k - z)$, and for the subsequent event the probability is $\mathrm{P}(y)$, where $y = z + X - k$. Page (1954) proposed that the probability of a test that starts from z can be generalised by the following Fredhorlm's integral equations of second kind given by

$$\mathrm{P}(z) = F(k - z) + \int_0^h \mathrm{P}(y) f(y + k - z)dy,\ 0 \leq z \leq h. \tag{3.21}$$

Similarly,

$$N(z) = 1 + \int_0^h N(y)f(X + k - z)dy, \; 0 \leq z \leq h \text{ and} \tag{3.22}$$

$$L(z) = 1 + L(0)F(k - z) + \int_0^h L(y)f(X + k - z)dy, \; 0 \leq z \leq h. \tag{3.23}$$

Page (1954) showed that the ARL of the CUSUM solution of equations (3.21) and (3.22) is given by

$$L(0) = \frac{N(0)}{1 - \mathrm{P}(0)}, \tag{3.24}$$

where $N(0)$ and $\mathrm{P}(0)$ are special cases where $z = 0$.

Equations (3.21) and (3.22) are usually solved numerically. Examples of such numerical results include the statistical nomograms for computing the ARL of the CUSUM by Woodall (1983) and Dobben de Bruyn (1968).

Siegmund (1985) approximation:

The Siegmund (1985) approximation formula for the ARL of the CUSUM is the most simple and the most widely used method. The formula is derived based on a one-sided CUSUM and it is given by

$$ARL = \frac{e^{-2\Delta b} + 2\Delta b - 1}{2\Delta^2}, \tag{3.25}$$

for $\Delta \neq 0$, $\Delta = \delta^* - k$ for the upper CUSUM and $\Delta = -\delta^* - k$ for the lower CUSUM, $k = (\mu_1 + \mu_0)/2$ is the reference point, $\delta^* = (\mu_1 - \mu_0)/\sigma$, σ is the standard deviation and $b = \log(B)$ is the upper boundary of the SPRT described in (3.2). The ARL under the null hypothesis is calculated when $\delta^* = 0$, and under the alternative hypothesis when

$\delta^* \neq 0$. The ARL of the two-sided CUSUM can be calculated from the formula

$$\frac{1}{L} = \frac{1}{L_1} + \frac{1}{L_2}, \tag{3.26}$$

where L_1 and L_2 are the ARLs of the lower and upper CUSUMs, and L is the ARL of the two-sided CUSUM.

Remark 3.1.2. *The ability of the CUSUM to include all information from the sequence of observations as well as detect small changes is important and makes it a desirable statistical tool for monitoring trends in meta-analysis. Unfortunately, in the CUSUM scheme it is required that the sequence of observations be independent and identically distributed, and this may not be satisfied in meta-analysis. However there are extensions of the CUSUM scheme developed by Gombay (2003), and Gombay and Serbian (2005) which have the CUSUM properties and can therefore serve as a good alternative procedure to monitor the trends in meta-analysis.*

3.1.3 Group sequential methods

Clinical trials investigate the effectiveness of new drugs or therapeutic procedures. Trials last for several weeks, months or years with their results accumulating continuously over the duration. Ethical, administrative as well as economic reasons often require that the accumulating data be evaluated at intervals to allow for early stoppage. Among the methods suggested for evaluation of the accumulating data at intervals is the use of repeated significance testing (see Armitage et al. (1969) and McPherson (1974)). However periodic evaluation of the accumulating data using standard significance testing can greatly inflate the Type I error (Armitage et al., 1969). ARMITAGE et al. (1975) showed that repeated significance testing can be a useful sequential method. Pocock (1977) used normal response with known variance to illustrate how the desired Type I error can be achieved in multiple significance testing of accumulating data. The procedures are referred to as the group sequential methods (see Chow et al. (2007), Jennison and Turnbull (2000)). They have multiple advantages. For example, group sequential methods are as efficient as the fully sequential methods in terms of low expected sample size and allow early stoppage of trials while retaining the overall error probability (Jennison and Turnbull, 2000). In meta-analysis they are used to address the issue of inflated Type I error in cumulative meta-analysis (see Pogue and Yusuf (1997), Higgins et al. (2011), Whitehead (1997a)) as well as to determine when sufficiency is attained in cumulative evidence.

To illustrate the general approach to group sequential methods, consider a clinical trial in which a treatment arm is being compared with control arm, and a planned total of N patients is divided into K groups. Let the response be a normal variable with variance σ^2, and the means μ_t and μ_c for treatment and control groups, respectively. The interest is to test the null hypothesis of no difference between the means, $H_0 : \theta = \mu_t - \mu_c = 0$ against $H_1 : \theta = \mu_t - \mu_c \neq 0$. Assume that equal number of n patients are accumulated in each arm of the experiment at each interim analysis. At the k-th interim analysis a standardized statistic is calculated as

$$Z_k = \frac{1}{\sqrt{2nk\sigma^2}} \left\{ \sum_{i=1}^{nk} X_{ti} - \sum_{i=1}^{nk} X_{ci} \right\}, \text{ for k=1, 2, ..., K,} \tag{3.27}$$

where X_{ti} and X_{ci} are the observations in the treatment and control groups, respectively. The variance σ^2 is unknown and is usually estimated from the sample data at each interim analysis. For each of the standardized statistics Z_k, a critical value C_k is chosen and the test terminates with the rejection of H_0 if $Z_k \geq C_k$; and if the test continues to the K-th analysis and $Z_K < C_K$, the test terminates and H_0 is accepted. In each of the analyses, a nominal level $\alpha^{'}$ is chosen to achieve a pre-specified Type I error probability α. In other words

$$\mathrm{P}(Z_1 \geq C_1 \text{ or } Z_2 \geq C_2 \text{ or...or } Z_K \geq C_K) = \alpha. \tag{3.28}$$

3.1.3.1 Sample size calculation based on power requirement

Sample size calculation based on power requirement is an important issue in planning controlled trials. Consider the problem of testing the null hypothesis of no treatment difference, $H_0 : \theta = \mu_t - \mu_c = 0$ against the two sided alternative $H_1 : \theta = \mu_t - \mu_c \neq 0$ with a significance level α and power $1 - \beta$ at $\mu_t - \mu_c = \pm\theta$. In a fixed sample size test the standardized statistic in (3.27) reduces to $Z_f = \frac{1}{\sqrt{2n\sigma^2}} \left\{ \sum_{i=1}^{n} X_{ti} - \sum_{i=1}^{n} X_{ci} \right\}$, and H_0 is rejected if $Z_f \geq z_{1-\alpha} + z_{1-\beta}$, where z is the critical value of the standard normal distribution (Jennison and Turnbull, 2000). The expected value

$$\begin{aligned} \mathrm{E}[Z_f] =& \mathrm{E}\left[\frac{1}{\sqrt{2n\sigma^2}} \left\{ \sum_{i=1}^{n} X_{ti} - \sum_{i=1}^{n} X_{ci} \right\} \right] \\ =& \pm\, \theta \sqrt{\{n/(2\sigma^2)\}}. \end{aligned} \tag{3.29}$$

Since it is often preferred to recommend a superior of the two treatments, the negative value is ignored in practice. Therefore, equating the positive value $\theta\sqrt{\{n/(2\sigma^2)\}}$ with

$z_{1-\alpha/2} + z_{1-\beta}$, and solving for n the sample size in the fixed sample test is given by

$$n_f = \left\{z_{1-\alpha/2} + z_{1-\beta}\right\}^2 2\sigma^2/\theta^2. \tag{3.30}$$

The maximum sample size n_g in a treatment arm of a group sequential test required to reject the null hypothesis with significance level α and power $1-\beta$ at $\mu_t - \mu_c = \pm\theta$ is a function of K, α and β, and is proportional to σ^2/θ^2 (Jennison and Turnbull, 2000). Since in the fixed sample test the sample size is also proportional to σ^2/θ^2, a ratio of the maximum sample size of group sequential test to the sample size of fixed sample test is defined as a function $R(K, \alpha, \beta)$. The values of $R(K, \alpha, \beta)$ are usually calculated numerically for different group sequential designs and provided as tables in many statistical textbooks. It follows that for a group sequential test with a maximum of K interim analyses the sample size per treatment arm n_g and the number of patients per treatment arm per group m_g needed to achieve a power requirement of $1-\beta$ are given by

$$n_g = R(K, \alpha, \beta)n_f \text{ and } m_g = R(K, \alpha, \beta)n_f/K, \tag{3.31}$$

respectively. For a detailed derivation and discussion see Jennison and Turnbull (2000), Chow et al. (2007) and the references therein.

3.1.3.2 Pocock's Test

The Pocock (1977) test is the most straightforward and widely used method. The Pocock's critical values, $C_k = C_P(K, \alpha)$ are functions of the Type I error probability, α and the total number of planned interim analyses K. In Pocock's test the repeated

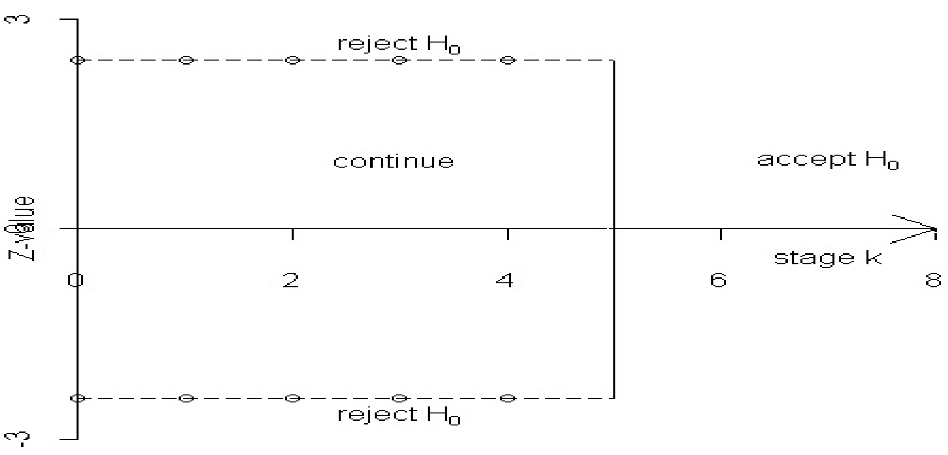


Figure 3.3: Pocock type monitoring boundaries with a maximum of 5 interim analyses

significance testing is conducted at a constant nominal level $\alpha^{'}$, and therefore the same critical value is used throughout the interim analyses. For example, the Pocock's critical values for two-sided test in a group sequential experiment with 5 interim analyses at 0.05 significance level are equal to $C_P(5, 0.05) = \pm 2.413$ (see Figure 3.3).

The decision in Pocock's test after analysis k=1, 2, ..., K-1 is that if $|Z_k| \geq C_P(K, \alpha)$ then stop and reject H_0; otherwise continue to analysis k+1. After analysis K, if $|Z_K| \geq C_P(K, \alpha)$ then stop, reject H_0; otherwise stop and accept H_0. The sample size per treatment arm and the number of patients per treatment per arm per group needed to achieve a power requirement of $1 - \beta$ in Pocock's test is determined in the same manner as in (3.31) with $R(K, \alpha, \beta) = R_P(K, \alpha, \beta)$.

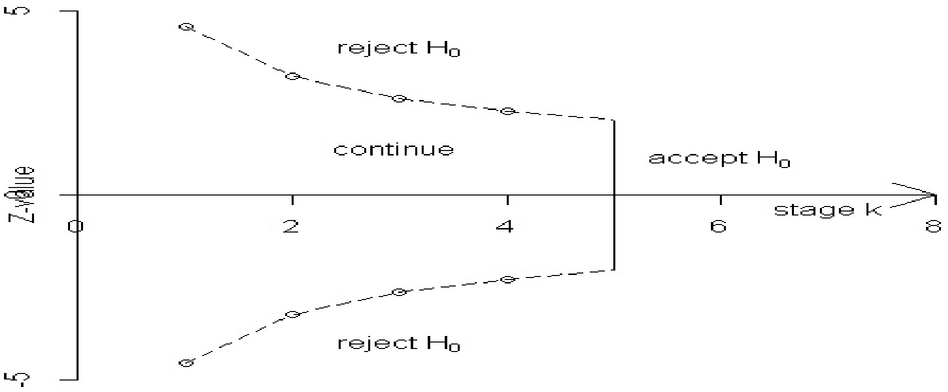


Figure 3.4: O'Brien-Fleming type monitoring boundaries with a maximum of 5 interim analyses

3.1.3.3 O'Brien-Fleming test

In the O'Brien and Fleming (1979) test the nominal significance level increases as the testing progresses, and therefore the test has relatively wider boundaries, and it is more conservative at the early stages (see Figure 3.4). This characteristic is desirable in clinical trials to prevent the possibility of spurious findings when information available in the analysis is still small. Another advantage of the O'Brien-Fleming test is that it allows the investigators to perform interim analyses at the last stage with a higher significance level nearly equal to the nominal level. The decision to continue or stop the trial is the same as in Pocock's test.

3.1.3.4 The error spending design

The Pocock's and O'Brien-Fleming tests require the number of interim analyses to be specified in advance and the groups be equally spaced with equal information size (number of observations). However, in practical situations these conditions are difficult to satisfy. For example, in consecutive clinical trials the decision to run another trial may lead to the choice of a larger or smaller sample size for the next study. As a result, the information accumulated at each interim analysis may not be equally spaced, and the implication is that the overall Type I error may be far from the target value (Chow et al., 2007).

Lan and DeMets (1983) introduced a flexible method that provides a solution to stopping boundary problem that can be readily adapted to clinical trials and cumulative meta-analyses (Pogue and Yusuf, 1997). The procedure is based on a spending function, $\alpha(t)$ which characterises the rate at which error rate is spent. The spending function, $\alpha(t)$ is non-decreasing. It assigns the proportion of the Type I error probability spent at each interim analysis. The variable t is the information fraction which at k-th interim analysis is determined by the total amount of information at k divided by the expected maximum information in the analysis, $t_k = I_k/I_{\max}$. The scale of information fraction, $t_{\max}$ is chosen so that the maximum is 1, and the spending function satisfies the conditions $\alpha(0) = 0$ and $\alpha(1) = \alpha$. The choice of a spending function, $\alpha(t)$ results in a choice of a particular group sequential method.

Suppose the standardized statistics Z_k for k=1, 2, ... in (3.27) correspond to the information fraction t_k and that $0 < t_1 < t_2 < ... < t_K = 1$. In the spending error design the critical value C_k at the information time t_k is determined from the accumulated information between 0 and t_k by solving the equation

$$\mathrm{P}(Z_1 \geq C_1 \text{ or } Z_2 \geq C_2 \text{ or...or } Z_k \geq C_k) = \alpha(t_k). \tag{3.32}$$

Note that the critical values C_k are determined by the spending function $\alpha(t)$ and the information fractions t_1, t_2, ..., t_k but does not depend on future information fractions or the number of interim analyses K. However if the experiment continues to the time t_{k+l}, the critical values C_k, C_{k+1}, ..., C_{k+l} can be defined to satisfy the probability $\mathrm{P}(Z_k < C_k \text{ or } Z_{k+1} < C_{k+1} \text{ or } Z_{k+2} < C_{k+2} \text{ or...or } Z_{k+l} \geq C_{k+l}) = \alpha(t_{k+l}) - \alpha(t_k)$, where $\alpha(t_{k+l}) - \alpha(t_k)$ is the increase in the significance level between the t_k and t_{k+l}. The alpha spending functions for Pocock and O'Brien-Fleming test are respectively given by

$$\alpha(t) = \min\left\{\alpha \log[1 + (e-1)t], \alpha\right\} \text{ and } \alpha(t) = 2\left\{1 - \Phi(z_{\alpha/2}/\sqrt{t})\right\}, \tag{3.33}$$

where Φ is the cumulative distribution function of the standard normal distribution.

3.1.3.5 Whitehead triangular test

The triangular test is described using the score statistics, $S_k = Z_k\sqrt{I_k}$ for k=1, 2, ..., K, where I_k is the information size corresponding to the k-th interim analysis. The score statistics S_k are assumed to be multivariate normal with $S_k \sim N(\theta I_k, I_k)$, and to have independent increments S_1, $S_2 - S_1$, ..., $S_K - S_{K-1}$ (Jennison and Turnbull, 2000). For group sequential testing of the null hypothesis $H_0 : \theta \leq 0$ against an alternative

$H_1 : \theta > 0$ with Type I error α at $\theta = 0$ and power $1 - \alpha$ at $\theta = \delta$, Whitehead and Stratton (1983) proposed a general continuation region for the score statistic defined as

$$S \in (L = -\tfrac{2}{\alpha} \log\left(\tfrac{1}{2\alpha}\right) + \tfrac{\delta}{4} I_k,\, U = \tfrac{2}{\delta} \log\left(\tfrac{3}{4\alpha}\right) + \tfrac{\delta}{2\alpha} I_k), \tag{3.34}$$

where I_k is the information at interim analysis k=1, 2, ..., K. These are the monitoring boundaries of the triangular test, and the maximum information of the test is chosen such that $0 < I_{\max} \leq \frac{8}{\delta^2 \log(\frac{1}{2\alpha})}$ (Jennison and Turnbull, 2000). If the boundaries meet at the final stage, the K-th interim analysis, then the maximum information of the triangular test can be calculated by

$$I_{\max} = \frac{4a}{\delta}. \tag{3.35}$$

For the case of unequal information increments at the interim analyses, Whitehead (1997b) used a result due to Siegmund (1985) and modified the boundaries of the triangular test in (3.34) to

$$L = -\tfrac{2}{\delta} \log\left(\tfrac{1}{2\alpha}\right) + 0.583\sqrt{\tfrac{I_{\max}}{K}} + \tfrac{3\delta}{4} \tfrac{k}{K} I_{\max} \text{ and } U = \tfrac{2}{\delta} \log\left(\tfrac{1}{2\alpha}\right) - 0.583\sqrt{\tfrac{I_{\max}}{K}} + \tfrac{\delta}{4} \tfrac{k}{K} I_{\max}. \tag{3.36}$$

The triangular test terminates at the first time when $S_k \notin (L, U)$, and at termination H_0 is rejected if $S_k > U$, and H_0 is accepted if $S_k < L$. Figure 3.5 is an example of a one-sided triangular test. The two-sided (double) triangular designs are also available.

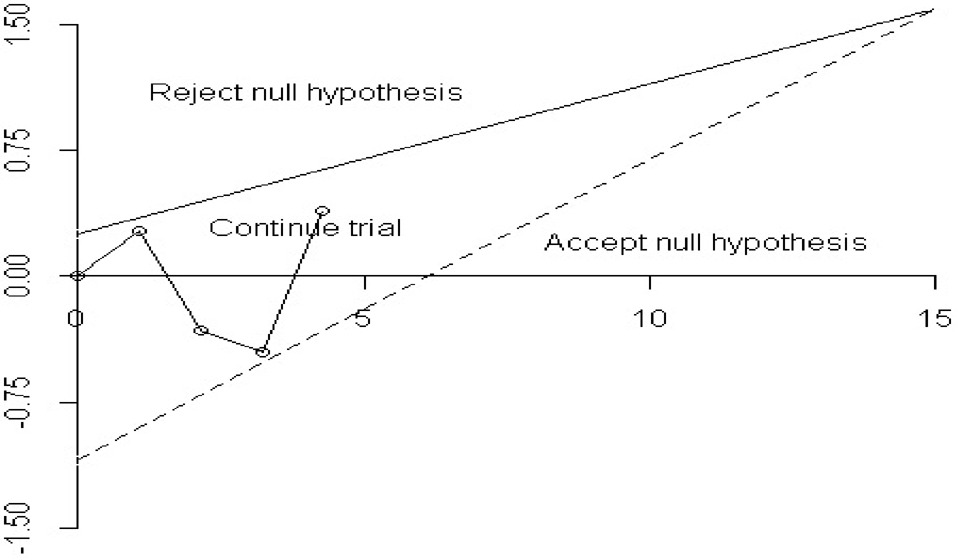


Figure 3.5: Whitehead triangular test

3.2 Methods for monitoring trends in meta-analysis

The methods for monitoring trends in meta-analysis can be subdivided into two groups: simplistic and sequential methods.

3.2.1 Simplistic methods for monitoring trends in meta-analysis

The "simplistic methods" for monitoring trends in magnitude of effect size can be described as initial or chronologically first approaches. They are easy and straightforward approaches in terms of calculation and interpretation of results, but have considerable draw-backs.

3.2.1.1 Homogeneity analysis

In meta-analysis, homogeneity analysis is usually conducted to test the hypothesis of no difference in treatment effects across studies. When testing the homogeneity of k studies, the Q statistic in (2.24) is compared with the chi-squared distribution with $k-1$ degrees of freedom, $Q = \sum\limits_{i=1}^{k} w_i(y_i - \hat{\theta}_{FEM})^2 \sim \chi^2_{k-1}$. If the null hypothesis is rejected, it means that the studies are heterogeneous. To use this method for monitoring temporal trends in meta-analysis, studies arranged in a chronological order are subdivided into subgroups according to publication year, for example, by decades, then homogeneity analysis is conducted across the subgroups. The approach is simple, and has been used by many researchers, see Higgins et al. (2003) as an example. However this method ignores the gradual character of temporal changes and their possible occurrence within as well as between study groups. Therefore, it is a 'crude method' (Kulinskaya and Koricheva, 2010).

3.2.1.2 Correlation/regression

Correlation and regression are popular statistical methods used to studying relationships between two or more variables. In meta-analysis, the relationship between the effect size and year of publication (Jennions and Møller, 2002) can be measured by the Pearson's product moment correlation given by

$$\rho(X, Y) = \frac{[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}, \tag{3.37}$$

where μ_X and μ_Y are means, and σ_X and σ_Y are standard deviations.

Alternatively, the trends in meta-analysis are often estimated using the regression slopes, see Shi and Copas (2004) for an example of a meta-regression model of alcohol use versus breast cancer. The regression model for monitoring changes in effects size in random-effects meta-analysis is based on the assumption given (Baker and Jackson, 2010) by

$$y_i \sim N(\theta + t_i\beta, \tau^2 + \sigma_i^2), \tag{3.38}$$

where y_i is the estimated treatment effect from the ith study at the time point t_i, θ is the mean treatment effect, β is the regression coefficient, τ^2 is the between-study variance and σ_i^2 is the within-study variance usually assumed to be known but estimated by the sampling variance. The regression model in (3.38) is valid only when linear trends are suspected, however Baker and Jackson (2010) proposed a general model for monitoring the trends that can accommodate both linear and non-linear trends based on the assumption given by $y_i \sim N(\theta \exp(-(T_n - T_i)\phi), \tau^2 + \sigma_i^2)$, where T_n is the date for the most recent study, T_i is the date for the ith study and ϕ is the shape parameter.

Correlation and regression constitute reasonable approaches for monitoring trends in meta-analysis, however the methods require that the magnitude of the effect sizes exhibit monotone increase or decrease with time which is not always possible (Kulinskaya and Koricheva, 2010, Leimu and Koricheva, 2004).

3.2.1.3 The use of standardized testing

Standardized testing may also be used to establish the presence or otherwise of temporal changes in a meta-analysis. For example, a null hypothesis of no difference between the results of a current study and the combined results of previous studies can be tested by calculating the Z statistic

$$Z = (y_k - \bar{y}_{k-1})\sqrt{w_k + \bar{w}_{k-1}}, \tag{3.39}$$

where y_k and $\bar{y}_{k-1}$ are the estimate from the current study and the combined effect of previous studies, respectively, and w_k and $\bar{w}_{k-1}$ are the corresponding weights, (see Ioannidis and Trikalinos (2005), Koricheva et al. (2013) for more discussion of this method).

3.2.2 Standard sequential methods for meta-analysis

Several sequential methods for meta-analysis have been proposed for monitoring temporal changes in magnitude of effect sizes. This Section reviews four different methods that are widely used.

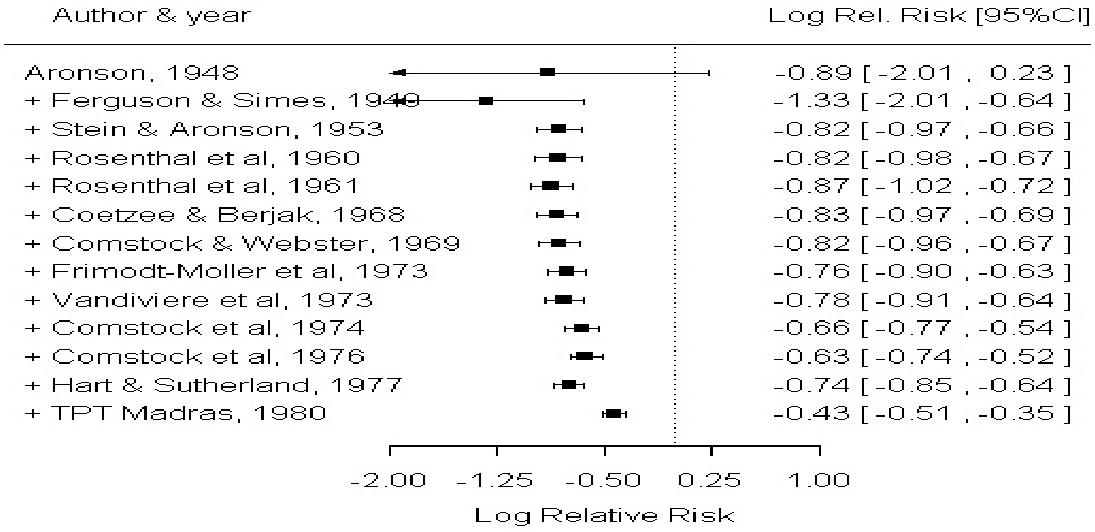


Figure 3.6: Forest plot of CMA of BGC vaccine for tuberculosis. Data and results are obtained from R library metafor (Viechtbauer, 2010), and results are combined using fixed effect model.

3.2.2.1 Cumulative meta-analysis

Historically, the first method proposed by Lau et al. (1992) was cumulative meta-analysis (CMA) which can be described as an open sequential test. The method involves pooling effect size estimates in a cumulative manner as new trial results are published. More exactly, CMA entails conducting a series of meta-analyses with successive addition of new effect size estimates from studies at interim analyses. Lau et al. (1992) had proposed the use of CMA for monitoring interventions across several randomized controlled trials, with the goal of understanding when evidence becomes definitive. CMA is routinely used for monitoring temporal changes in effect sizes (see Lau et al. (1992),

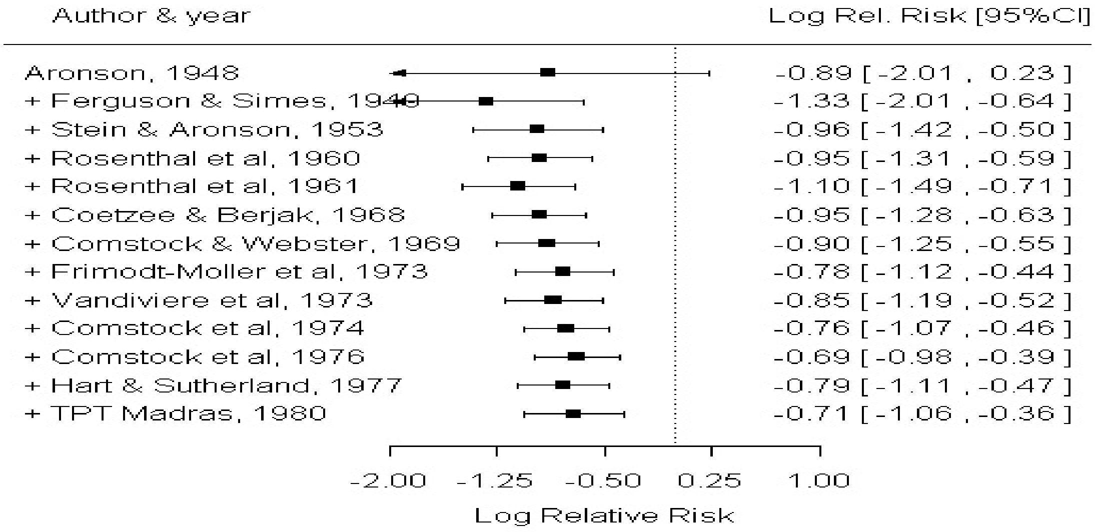


Figure 3.7: Forest plot of CMA of BGC vaccine for tuberculosis. Data and results obtained from R library metafor (Viechtbauer, 2010), and results are combined using random-effects model.

Ioannidis and Trikalinos (2005), Leimu and Koricheva (2004)). When the results are arranged in a chronological sequence according to year of publication, the plotted values of the combined effects, $\hat{\theta}_k$ and confidence intervals calculated consecutively for k=1, 2, ..., K can reveal temporal patterns (Kulinskaya and Koricheva, 2010).

Suppose y_i for i=1, 2, ... are the effect size estimates obtained sequentially. The cumulative effect at the k-th interim analysis is estimated by $\hat{\theta}_k = \sum\limits_{i=1}^{k} w_i y_i / \sum\limits_{i=1}^{k} w_i$, where w_i are the weights assigned to the studies according to the meta-analytic model used. Plotted values of estimates of cumulative effects against time allow visual monitoring of tangential increase or decrease in effect size over time. Figures 3.6 and 3.7 are examples of forest plots of CMA of BCG vaccine for tuberculosis. Data and results are obtained from R library metafor (Viechtbauer, 2010) and results are combined using fixed- and random-effects models, respectively.

Cumulative meta-analysis can reveal whether there is consistency in the results of consecutive studies and indicate a point at which no further studies are required because the definitive conclusion is reached. Further, it has the advantage of revealing uneven irregular changes in effect sizes as well as multiple shifts in opposite directions (Leimu and Koricheva, 2004). CMA as a graphical tool is useful for initial inspection of the data, but as in any visual method it might be subject to misinterpretation, and therefore needs to be supplemented by a formal statistical method (Kulinskaya and Koricheva, 2010, Leimu and Koricheva, 2004, Koricheva et al., 2013). In addition, by definition CMA involves repeated analysis of the accumulating evidence and thus, even if there is no treatment effect, multiple testing involved leads to the inflation of Type I error.

3.2.2.2 Sequential meta-analysis

The second group of methods is sequential meta-analysis (SMA). These methods involve the use of formal group sequential boundaries to monitor CMA and were proposed by Pogue and Yusuf (1997) to address the issue of inflated Type I error in CMA. The crossing monitoring boundaries of group sequential methods can indicate significant change in cumulative effect and may be used to stop a meta-analysis when there is sufficient evidence of effect based on pre-specified significance level and power (Higgins et al., 2011). There are several group sequential designs which can maintain the overall significance level, however Lan and DeMets (1983) alpha spending method is more flexible in the sense that it does not require the number or the times of interim analyses to be specified in advance (See Section 3.1.3.4). Since meta-analysis is a continuous process and the number of interim analyses is not known in advance, Pogue and Yusuf (1997) suggested the use of Lan and DeMets (1983) method for SMA.

A key issue in conducting a SMA is the calculation of the optimum information size (OIS) needed to define the monitoring boundaries. The OIS is the amount of information needed to detect a significant treatment effect had a well-designed trial been planned. It is a function of the maximum sample size required to achieve the power requirement for the test at the given significant level. Lan and DeMets (1983) used standard methods with small significance level α and high power $1-\beta$ of 90% or 95% to calculate the maximum sample size. For example, for the mortality rates P_c of 10% in the control group, and treatment effect, $\Delta = P_c - P_t$, the Type I error may be set at 1% and the power at 90%. The sample size per treatment required to achieve the power requirement is given by

$$n = 2 \times \frac{(\bar{Z}_{\alpha}+\bar{Z}_{1-\beta})^2/2P^*(1-P^*)}{\Delta^2}, \text{ where } P^* = (P_c + P_t)/2 \tag{3.40}$$

The calculation of the OIS is based on fixed effects model and hence the method is only appropriate for FEM. A number of methods were proposed to correct this. Wetterslev et al. (2008) used a heterogeneity inflated OIS to account for heterogeneity in treatment effects, but this method is problematic (Kulinskaya and Wood, 2014). Whitehead (1997a) describes the use of standard stopping boundaries for random-effects meta-analysis. Bollen et al. (2006) and van der Tweel and Bollen (2010) used the double triangular test in a retrospective meta-analysis. Higgins et al. (2011) proposed a sequential method for random-effects meta-analysis that uses a semi-Bayes procedure to update evidence on the among-study variance, starting with an informative prior distribution that may be based on findings from a previous meta-analyses. A common issue for these methods is that the monitoring boundaries are generally defined based on FEM and do not incorporate the presence of heterogeneity in treatment effects. As a result, as revealed by simulations, these methods have shown a considerable inflation of the Type I error when the values of τ^2 are large, see Higgins et al. (2011), Wetterslev et al. (2008). Therefore using such methods in random-effects model can lead to spurious statistical inference.

3.2.2.3 Use of quality control charts

In the theory of control charts, variability in on-line process measurements is assessed by constructing monitoring boundaries. These boundaries are also known as control limits and are constructed based on the distribution of the observed values of the process. When the process mean is within the control limits, the process is said to be statistically in-control (variability is due to chance). However if at any stage the process mean crosses the control limits, the process will be considered to be out-of-control (variability is due to assignable causes and corrective action needs to be taken). There are several quality control charts including Shewhart (1931) and Page (1954) CUSUM charts, see Section 3.1.

Kulinskaya and Koricheva (2010) proposed the use of quality control charts for detection of outliers and temporal trends in meta-analysis. The use of QC charts in meta-analysis is straightforward if the sequential effect estimates are independent and their distribution can be approximated by the normal distribution. For example, let $y_i \sim N(\theta, \sigma_i^2)$ for i=1, 2, ... be estimates of effect size from consecutive studies and θ be a target value. The control limits for monitoring the meta-analysis are determined based on the values of θ and σ_i^2. As more studies are conducted and results are combined, if the mean of the process is close to θ, the process is said to be in-control (no change in treatment effect). However, when adding the results of new studies and the mean crosses the control limits, then the process will be considered to be out-of-control (there exists a change in treatment effect). The method is simple and had successfully been applied to fixed effect model. However for random-effects model the estimation of τ^2 introduces dependency between the sequential effects (Kulinskaya and Koricheva, 2010) and hence their distribution is not consistent with the standard assumptions of the QC charts.

3.2.2.4 Penalised Z testing

The last group of methods involves the "penalised Z test" introduced by Lan et al. (2003). This is an alternative approach to address the issue of inflated Type I error in CMA. The method is based on the use of the law of iterated logarithm to 'penalize' for the multiple testing in CMA. The usual Wald test for significance of the combined effect at the k-th interim analysis is adjusted by a constant factor, and is defined by

$$Z^*(k) = \frac{S(k)}{\sqrt{\lambda \Gamma_k \log\log(\Gamma_k)}}, \tag{3.41}$$

where λ is the adjustment factor determined using simulation, $S(k)$ is the sum of the estimates of treatment effects up to the k-th interim analysis and Γ_k is the sum of weights assigned to studies. Lan et al. (2003), Hu et al. (2007) claim that the 'penalized Z test' exhibits a good control of the Type I error in CMA both in FEM and REM when a reasonable value of λ is used. For example, the value of $\lambda = 1.5$ was found to control the Type I error in FEM, while the value of $\lambda = 2$ was found to control the Type I error in REM when relative risks, odds ratio and risks difference effect sizes were used to combine results of up to 25 studies (Hu et al., 2007). The choice of λ is important in controlling the Type I error, however its value varies according to the type of effect measure, number of studies, average studies size and amount of heterogeneity in the treatment effects. Therefore the determination of the 'reasonable value of λ' can be difficult in practice.

3.3 Adaptive clinical trials

In clinical trials, trial procedures and statistical methods are usually pre-specified at the beginning of the trial. However if they are wrongly chosen this may lead to failure in the study. Adaptive clinical trials allow modification of trial procedures and the statistical methods in an ongoing trial based on data accumulating during the progress of the trial, while maintaining the integrity and validity of the trial. The purpose is to provide investigators the flexibility to identify the best/optimal clinical benefit of the test treatment under study without compromising the validity and integrity of the intended study (Chow et al., 2008), as well as increase the probability of success of the intended trial. The procedure in adaptive clinical trials is such that at some stages during the trial a decision is made on whether to abandon or continue with the trial, based on review of accumulated information from preceding stages. If the decision is to continue, then the next study is designed using the results from the previous stages (Jennison and Turnbull, 2005). There are several adaptive methods in clinical trials that include sample size re-estimation, adaptive randomization, adaptive dose finding and adaptive hypotheses, but for the purpose of this work only sample size re-estimation is discussed here.

3.3.1 Sample size re-estimation

In calculating the sample size of a clinical trial, investigators usually make assumptions about the expected treatment effect and the variance of the outcome variable(s). If the assumptions are not correct and the actual values of the parameters differ substantially from the expected, the sample size may be too small or large, thereby under-powering or overpowering the study, both of which have serious consequences. For example, an underpowered study may lead to inability to detect the treatment effect and making the trial inconclusive. While an overpowered study leads to wastage of resources that may be used elsewhere. Sample size re-estimation allows the parameter estimates to be updated during an ongoing trial, and then used to modify the sample size accordingly.

Suppose that after n_1 observations per group have been taken, the standardized statistic $Z_1 = \sqrt{n_1/2\hat{\sigma}^2}(\bar{Y}_1 - \bar{X}_1)$ is computed. Assume n_1 is large enough so that

σ_2 can be estimated after n_1 observations has been taken. Proschan and Hunsberger (1995) proposed a standard method for calculating the number of the additional observations $n_2 = n_2(z_1)$ based on the observed z_1 using conditional power (CP). Let $CP_\delta(n_2, z_\alpha|z_1)$ denote the conditional probability that Z based on $n_1 + n_2$ observations exceeds the critical value z_α, given that $Z_1 = z_1$, and that $(\mu_y - \mu_x)/\sigma = \delta$. Proschan and Hunsberger (1995) conditional power approach is given by

$$\begin{aligned} CP_{n_2,Z_\alpha|Z_1} =& \Pr\left(Z > Z_\alpha | Z_1 = z_1, \delta\right) \\ =& 1 - Z\left\{\frac{Z_\alpha\sqrt{2(n_1+n_2)} - Z_1\sqrt{2n_1} - n_2\delta}{\sqrt{2n_2}}\right\}, \end{aligned} \tag{3.42}$$

where the treatment difference in standardized form δ is replaced with the observed estimate.

Sample size calculation generally requires that the standard deviation of the process/observations be known. When the standard deviation is unknown, it is estimated from previous data on the topic or at least from a pilot study. However treating estimates obtained from data observed at interim stage as true values lead to the same problem faced at the original calculation of the sample size before conducting the study. More so, when a clinical trial starts with a small sample size, re-estimating the sample size based on observed treatment difference instead of the actual clinical difference that need to be detected can cause bias and be misleading (Chow et al., 2008). Chapter 6 provided more discussions on sample size re-estimation with regard to sequential bias in accumulating evidence in meta-analysis.

Chapter 4

A sequential method for random-effects meta-analysis

As earlier mentioned in the introductory Chapter, temporal changes in magnitude of effect sizes reported in many areas of research can impair the validity of results and conclusions in meta-analysis. Standard sequential methods proposed for monitoring the trends are based on solid statistical theory only in fixed effect approach, and therefore are not suitable for sequential random-effects meta-analysis. The major obstacle in simple application of the sequential methods in random-effect meta-analysis is handling of the between-study variance, τ^2 (Higgins et al., 2011). When the number of studies are few many estimators of τ^2 underestimate it. This Chapter introduces the use of a truncated CUSUM-type test (Gombay method) for the sequential random-effects meta-analysis. The Gombay method is a sequential change detection test for parametric models in the presence of a nuisance parameter. A nuisance parameter is a parameter that is not of immediate interest but must be accounted for in the course of the analysis. So in the application of the method in random-effects meta-analysis, the between-study variance is treated as the nuisance parameter. The Gombay method is a large-sample method suitable for most practical problems whether their observations are independent or not.

Section 4.1 describes the Gombay method. Section 4.2 formulates the Gombay method for random-effects model of meta-analysis. Section 4.3 is the report on a simulation study of the Gombay test for sequential REM with standard critical values obtained from asymptotic theory. Discussion of the Gombay method based on critical values derived from asymptotic theory is presented in Section 4.4.

4.1 The Gombay Method

Before presenting the Gombay method we briefly introduce the score test which is closely related to the Gombay method.

4.1.1 Score test

The score test introduced by Rao (1948) is a fixed sample size test of a null hypothesis that a parameter of interest takes a particular value. Let X be an independent random variable with density $f(X, \omega)$, where ω is a parameter of interest. The score test statistic for testing a null hypothesis $H_0 : \omega = \omega_0$ is defined (Rao, 1948) by

$$S(\omega) = \frac{[\mu(\omega)]^2}{I(\omega)}, \tag{4.1}$$

where $\mu(\omega) = \frac{\partial}{\partial\omega}\log\left[f(X,\omega)\right]$ is known as the score vector and $I(\omega) = \text{var}\left[\mu(\omega)\right] = \text{E}_X\left[(\mu(\omega))^2\right] = \text{E}_X\left[\left(\frac{\partial}{\partial\omega}\log[f(X,\omega)]\right)^2\right]$ is the Fisher information and the derivatives taken at ω_0. Under the null hypothesis the statistic $S(\omega)$ is χ^2 distributed with 1 degrees of freedom (Rao, 1948).

In most statistical problems ω is rarely only a parameter of interest. Let $\omega = (\theta, \eta)$, such that the observed variable $X \sim f(., \theta, \eta)$, θ is a vector of real parameters of interest and η is the vector of nuisance parameter. Since the interest is in inference about the parameters of interest θ, it is important to find a way to deal with the nuisance parameter. One way to eliminate the nuisance parameter is by conditioning the score statistic, (see (Lindsey, 1983, Basu, 1977)). A suitable statistic is chosen, say $g(x \in X) : (x \in X, \omega) \to (y \in Y, \theta)$ such that the conditional distribution of μ^c depends on ω only through θ. The conditional score vector may be defined (Lindsey, 1983) by

$$g(x \in X) = \mu(\omega) - \text{E}\left[\mu | T\right], \tag{4.2}$$

where T is a sufficient statistic whose sampling distribution depends on θ only. If θ is real-valued, the information corresponding to $g(x \in X)$ is obtained from a Fisher information matrix for the parameters (θ, η) given (Gombay and Serbian, 2005, Lindsey, 1983) by

$$I = \begin{pmatrix} I_{\theta\theta} & I_{\theta\eta} \\ I_{\eta\theta} & I_{\eta\eta} \end{pmatrix},$$

and the marginal information about θ, also known as the effective information, is given by $I(\theta) = I_{\theta\theta} - I_{\theta\eta}I_{\eta\eta}^{-1}I_{\eta\theta}$, see (Bera and Bilias, 2001, Gombay and Serbian, 2005).

4.1.2 Sequential hypotheses and Gombay test statistic

The Gombay method is a sequential change detection test for parametric models in the presence of a vector nuisance parameter. It is closely related to the score test described above. However the score test is a fixed sample size test of a null hypothesis that a parameter of interest takes a particular value, while the Gombay test is a sequential change detection test with the test statistic defined by the maximum of a sequence of score statistics $S_j = S|X_{(j)} = c(X_1, X_2, ..., X_j)$ calculated from the sequence of observed data, $G_k = \max\{S_1, S_2, ..., S_k\}$. Below we describe the Gombay method introduced as test I in Gombay and Serbian (2005). Consider a sequence of independent random variables (r.v.) X_1, X_2, ... $\sim f_{\theta_i,\eta_i}$, where f is a probability density function, θ is a (vector) parameter of interest and η is a nuisance parameter. Consider a test for the composite hypothesis

H_0: $\theta_i = \theta_0$, $\eta_i = \eta$; $i = 1$, 2, ... against alternatives

$$H_1\text{:}\ \begin{cases} \theta_i = \theta_0, \eta_i = \eta; & i = 1, 2, ...r, \\ \\ \theta_i = \theta_0 + \Delta\theta, \eta_i = \eta; & i \geq r+1, \end{cases}$$

where $r \geq 1$ is an unknown time of change, $\Delta\theta$, a shift in the value of the parameter of interest from θ_0 and η an unknown nuisance parameter. The null value of the vector of parameter of interest θ_0 can take any value from R^d. In the context of meta-analysis, comparing a treatment and a control group, the null value of the effect parameter may be set at $\theta_0 = \theta_T - \theta_C = 0$; while for a meta-analysis of stage IV clinical trials with the research interest to detect any possible shift from a known effect of a treatment,

Table 4.1: Critical values of two-sided Gombay test $C(\alpha)$, χ_1^2 distribution $\chi_1^2(\alpha)$ and standard normal distribution $Z_{1-\alpha}$

K	10	10	50	50	50	100	100	100	1000	1000	1000
α	0.025	0.05	0.010	0.025	0.05	0.010	0.025	0.05	0.010	0.025	0.05
$C(\alpha)$	4.0177	3.4710	4.5032	3.9438	3.5164	4.4892	3.9606	3.5566	4.5062	4.0363	3.6772
$C^*(\alpha)$	4.5544	4.0077	4.9229	4.3635	3.9360	4.8859	4.3572	3.9532	4.8588	4.3889	4.0297
$Z_{1-\alpha}$	1.9599	1.6449	2.3263	1.9599	1.6449	2.3263	1.9599	1.6449	2.3263	1.9599	1.6449

the null hypothesis may be set at a value other than zero, say $\theta_0 = \theta_p$, where θ_p is the value of the known effect of the treatment from results of previous studies.

Denote $\psi = (\theta, \eta)$. The log-likelihood function at the k-th interim analysis is $l(\psi) = \sum_{i=1}^{k} \ln f(X_i, \psi)$, and the score vector for θ and η is defined by

$$V_k(\theta_0, \eta) = \frac{\partial l(\psi)}{\partial \psi} = \sum_{i=1}^{k} \frac{\partial}{\partial \psi} \log f_{\theta_0 \eta}(X_i). \tag{4.3}$$

In order to define a test statistic for the hypotheses about θ, a Fisher information matrix I for k observations is partitioned as

$$I = \begin{pmatrix} I_{\theta\theta} & I_{\theta\eta} \\ I_{\eta\theta} & I_{\eta\eta} \end{pmatrix},$$

where

$$I_{11} = \left(-\mathrm{E}\tfrac{\partial^2}{\partial\theta^2} l(\theta, \eta)\right),\ I_{22} = \left(-\mathrm{E}\tfrac{\partial^2}{\partial\eta^2} l(\theta, \eta)\right) \text{ and } I_{12} = I_{21}^t = \left(-\mathrm{E}\tfrac{\partial^2}{\partial\theta\partial\eta} l(\theta, \eta)\right).$$

Replacing the nuisance parameter η with its restricted maximum likelihood estimate $\hat{\eta}_k$, obtained from the solution of

$$\sum_{i=1}^{k} \frac{\partial}{\partial \eta} \log f(X_i : \theta_0, \eta) = 0, \tag{4.4}$$

the conditional efficient score vector V_k is given by

$$V_k(\theta_0, \hat{\eta}_k) = \sum_{i=1}^{k} \frac{\partial}{\partial \theta} \log f_{\theta_0 \hat{\eta}}(X_i). \tag{4.5}$$

This vector is also sometimes termed effective score vector and its variance $\Gamma_k(\theta_0, \eta) = I_{11} - I_{12} I_{22}^{-1} I_{21}$ is called effective information, (Bera and Bilias, 2001). Note that for independent and identically distributed r.v.'s, this variance increases linearly with the number of observations: $\Gamma_k(\theta_0, \eta) = k\Gamma_1(\theta_0, \eta)$. Under some standard regularity conditions guaranteeing the existence and consistence of a sequence of maximum likelihood estimates given by Serfling (1980) and Lehmann (2001), Gombay and Serbian (2005) showed that under H_0, as $k \to \infty$, the effective score vector can be written as

$$\begin{aligned} V_k(\theta_0, \hat{\eta}_k) =& \sum_{i=1}^{k} \frac{\partial}{\partial \theta} \log f_{\theta \hat{\eta}_k} \\ =& \sum_{i=1}^{k} \left\{ \frac{\partial}{\partial \theta} \log f_{\theta_0 \eta} \right\} \\ & - \sum_{i=1}^{k} \left\{ \frac{\partial}{\partial \eta} \left(\log f_{\theta_0 \eta} \right) I_{22}^{-1}(\theta_0, \eta) I_{21}(\theta_0, \eta) \right\} \\ & + O(\log \log k) \\ =& \sum_{i=1}^{k} Z_i + O(\log \log k), \end{aligned} \tag{4.6}$$

where Z_i are independent identically distributed (i.i.d.) random variables with expected value $\mathrm{E}[Z_i] = 0$ and the covariance matrix $\mathrm{cov}(Z_i) = k^{-1}\Gamma_k(\theta_0, \eta)$, for $\Gamma_k(\theta_0, \eta) =$

$I_{11} - I_{12}I_{22}^{-1}I_{21}$. It follows that the statistic

$$T_k = \sqrt{k}\Gamma_k(\theta_0, \eta)^{-1/2} \sum_{i=1}^{k} \frac{\partial}{\partial\theta} \log f_{\theta_0, \hat{\eta}_k} \tag{4.7}$$

is asymptotically ($k \to \infty$) the sum of i.i.d. random variables with mean 0 and variance 1, and thus a sequence of statistics $\{T_k\}$ can be approximated by a standard Wiener process. In order to use the statistic T_k, for testing the covariance $\Gamma_k(\theta_0, \eta)$ is replaced with its estimate $\Gamma_k(\theta_0, \hat{\eta}_k)$. Gombay (2003) and Gombay and Serbian (2005) introduced a sequential change detection test based on statistic T_k in (4.7) as follows. For $k = 2, 3, \cdots, K$, where K is a truncation point, reject H_0 in favor of a positive change $\Delta\theta > 0$ at time k if

$$G(K) =^{\max}_{1<k\leq K} \frac{1}{\sqrt{K}} T_k \geq \sqrt{K} C(\alpha) \tag{4.8}$$

and if no such k, $k \leq K$, exists do not reject H_0. The asymptotic critical values $C_(\alpha)$ of this 1-sided test are calculated by

$$C(\alpha) =(2 \log\log K)^{-\frac{1}{2}}(-\log(-\log(1-\alpha)) + 2.5 \log\log K - \frac{1}{2}\log\pi), \tag{4.9}$$

where α is the significance level and K is the truncation point or the maximum number of observations. For the two-sided test based on $|T_k|$, the critical values are given by

$$C^*(\alpha) =(2 \log\log K)^{-\frac{1}{2}}(-\log(-\frac{1}{2}\log(1-\alpha)) + 2.5 \log\log K - \frac{1}{2}\log\pi). \tag{4.10}$$

Table 4.1 shows that the critical values of the Gombay test decrease with increase in maximum number of observations (studies) K, and with increase in the value of significance level α. The critical values of the Gombay test are higher compared with the critical values $Z_{1-\alpha}$ obtained from standard normal distribution. See Gombay

(2003) and Gombay and Serbian (2005) for a detailed derivation and discussion of this method.

4.2 Formulation of the Gombay test statistic for REM

To apply the Gombay method in random-effects model of meta-analysis, assume studies are conducted independently and sequentially over time. However, in practice a difficulty can arise in determining the order in which studies are performed or published. For example the year of publication of two or more studies may coincide. Where such difficulty arises the order is selected randomly. Each study estimates a treatment effect, y_i for i=1, 2, ... with variance σ_i^2. Assume that there is no correlation between the effect size estimates and the variances. Under the null hypothesis, H_0, each effect estimate is normally distributed with the same mean θ, $y_i \sim N\left(\theta, (\hat{w}_i^*)^{-1}\right)$, where $\hat{w}_i^* = (\hat{\tau}^2 + \sigma_i^2)^{-1}$ is the estimate of the weight in random effects model. The mean parameter, θ is the population treatment effect and it is estimated at the step k as weighted mean of the individual effect estimates, $\hat{\theta}_k = \sum_{i=1}^{k} \hat{w}_i^* y_i / \sum_{i=1}^{k} \hat{w}_i^*$, k=1, 2, ... Let $\theta = \theta_0$ be the null value of the effect parameter. As more studies are conducted and results are continually combined, the goal is to determine when the combined effect, $\hat{\theta}_k$ changes significantly from the null value, θ_0 if at all, and stop further studies.

The log likelihood function required to define the Gombay test statistic is given by

$$L\left(y_i : \theta, \tau^2\right) = \frac{1}{2}\left\{ \log \hat{w}_i^* - \hat{w}_i^*(y_i - \theta_0)^2 + C \right\}, \tag{4.11}$$

where C is a constant. The efficient score statistic is $V_k = \sum_1^k \hat{w}_i^*(y_i - \theta_0)$. This familiar statistic is routinely used in meta-analysis for testing a value of the mean in k studies. Its variance is $\Gamma_k = \sum_1^k \mathrm{E}[\hat{w}_i^*]$. In the sequential setting, the Gombay test statistic is based on the maximum of the standardised and scaled by $\sqrt{k}$ score statistics (4.8) given by

$$T_k = \frac{\sqrt{k}\sum\limits_{i=1}^{k} \hat{w}_i^*(y_i - \theta_0)}{\sqrt{\sum\limits_{i=1}^{k} \mathrm{E}[\hat{w}_i^*]}}. \tag{4.12}$$

Because the probability distribution of $\hat{\tau}^2$ is unknown, the expected value of the estimated weight $\hat{w}_i^*$ in (4.12) needs to be approximated. Assuming that the expected value $\mathrm{E}[\hat{\tau}_i^2] = \tau^2$ for i=1, 2, ..., K, the expected value of the estimated weight estimates in (4.12) can be approximated by the first term in their Taylor series expansion, $\mathrm{E}[\hat{w}_i^*] = w_i^*(\tau^2)$. The between-study variance component τ^2 is estimated using the full information available from k studies, $\hat{\tau}_k$, or from all K studies, $\hat{\tau}_K^2$.

We proposed a sequential test using the weights $w_i^* = w_i(\hat{\tau}_k^2)$ and $\mathrm{E}[\hat{w}_i^*] = w_i^*(\hat{\tau}_k^2)$ in (4.12), and based on the maximum (over all $k \leq K$) of $\sqrt{k}T_k$, see Dogo et al. (2015). The τ^2 was estimated by one of the methods by DerSimonian and Laird (1986); Higgins et al. (2011); Paule and Mandel (1982) and the REML. In what follows, the Gombay test statistics based on the four above estimators are denoted by GDL, GH, GMP and $GREML$, respectively.

4.3 The simulation study.

The objectives of the simulation study presented in this Section is to evaluate the Type I error rate and the power of Gombay test for REM with standard critical values in relation to the number of studies K in the meta-analysis, average studies sizes n, and the amount of heterogeneity in the treatment effects τ^2. The simulation study also compares the performance of the Gombay test for REM with standard critical values based on four different estimators of τ^2; DerSimonian and Laird (1986), Higgins et al. (2011), Paule and Mandel (1982) and the REML.

The data for the simulations were generated as follows. For studies i=1, 2, ..., K, the sample sizes were generated from normal distribution, $n_i \sim N\left(n, \frac{n}{4}\right)$ rounded to the nearest integer and values less than 3 were truncated at 3. The sample variances, S_i^2 were generated from the scaled Chi-squared distribution, $S_i^2 \sim \frac{\sigma_i^2}{(n_i-1)}\chi^2_{n_i-1}$. The effect size estimates were generated from normal distribution, $y_i \sim N\left(\Delta\theta, \sigma^2/n_i + \tau^2\right)$, where $\Delta\theta$ is the difference in the null value of effect parameter θ_0 and the alternative $\theta_0 + h$. Critical values were calculated based on 5 % significance level and the null value of the effect parameter set at $\theta_0 = 0$. The sequential testing starts with a minimum of two studies and stops as soon as a boundary value is reached or after the K^{th} interim analysis. For each combination of the following variables: $\sigma^2 = 1$, $\Delta\theta = (0.00, 0.05, 0.10, 0.15, 0.20)$, $n = (20, 50, 100, 1000)$, $K = (10, 30, 50)$ and $\tau^2 = (0.00, 0.015, 0.030, 0.045, 0.060)$, a total of 10,000 simulations were conducted, then the empirical power of the test to reject H_0 was calculated and recorded. The values

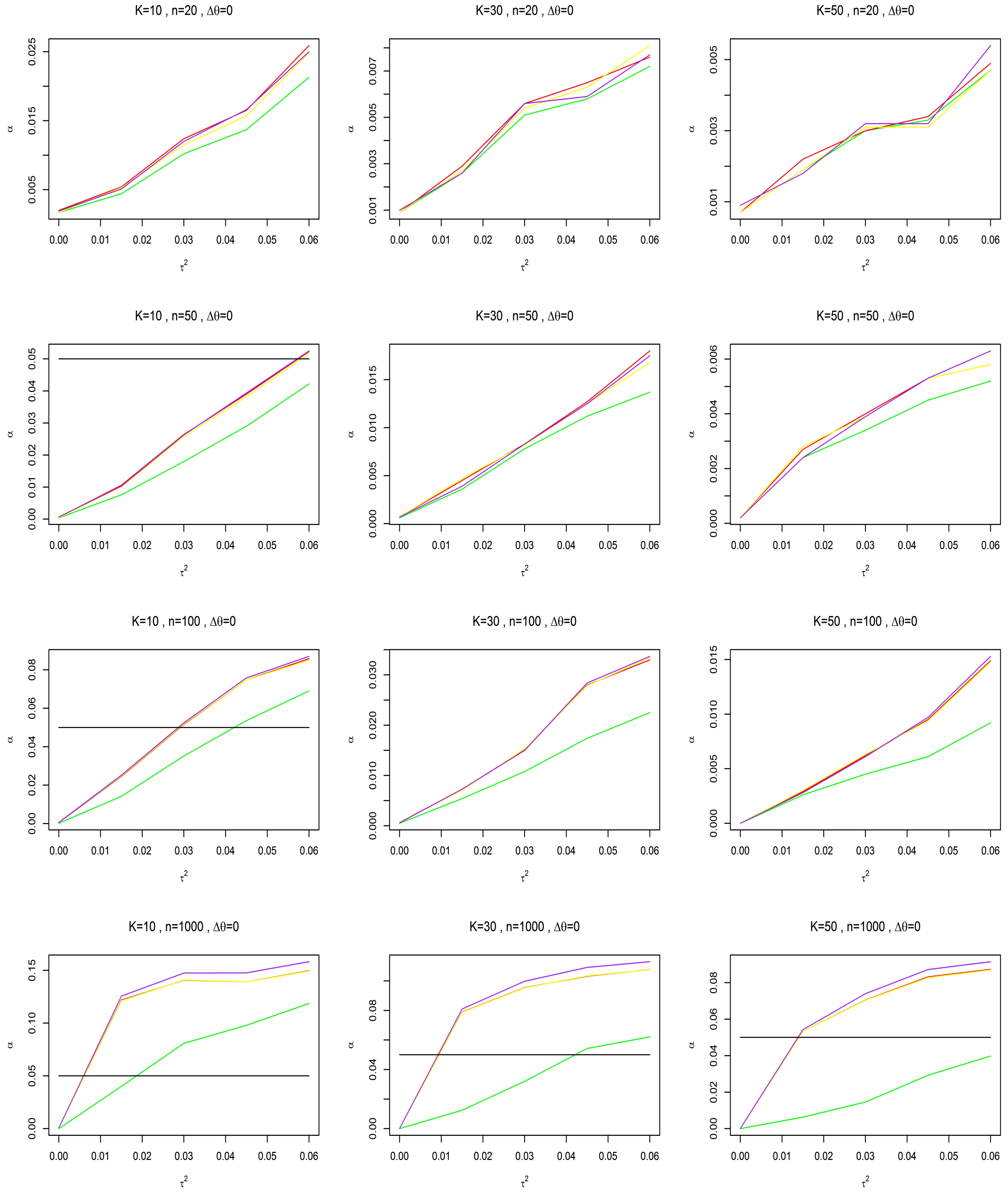


Figure 4.1: Overall Type I error achieved by Gombay test for REM with standard critical values at the nominal 0.05 level based on DerSimonian and Laird (1986); Higgins et al. (2011); Paule and Mandel (1982) and REML estimators of τ^2 (GDL -red line, GH - green, GPM - yellow and GREML - purple line, respectively). K is the number of studies; n is the average sample size of studies; $\Delta\theta$ is the amount of shift in value of the effect parameter from θ_0, τ^2 is the value of the between-study variance. The black straight line represents the nominal 0.05% level for the test.

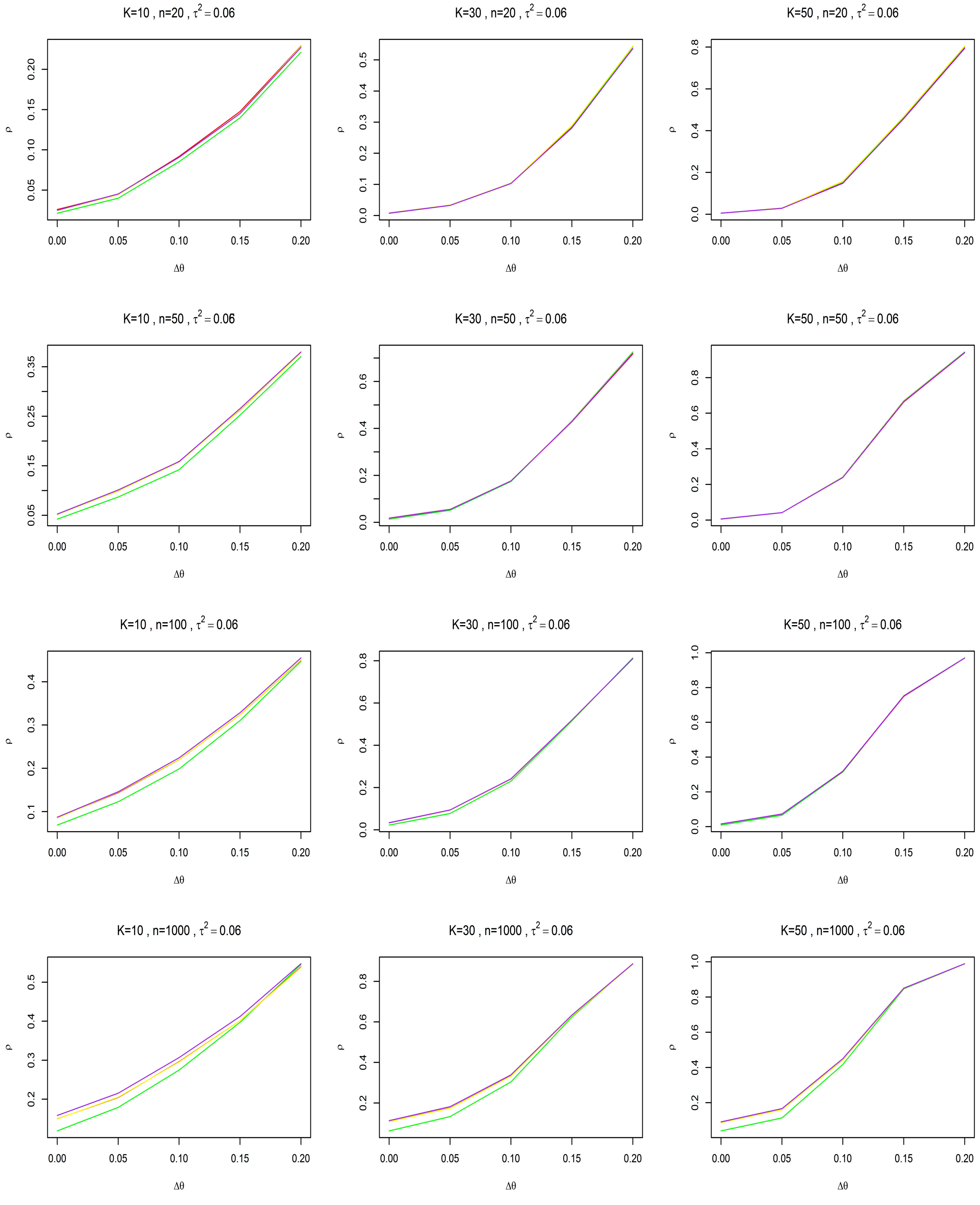


Figure 4.2: Power of Gombay test for REM with standard critical values at the nominal 0.05 level based on DerSimonian and Laird (1986); Higgins et al. (2011); Paule and Mandel (1982) and REML estimators of τ^2 (GDL -red line, GH - green, GPM - yellow and GREML - purple line, respectively). K is the number of studies; n is the average sample size of studies; ρ is the power of the test and $\Delta\theta$ is the amount of shift in value of effect parameter from θ_0; τ^2 is the between-study variance.

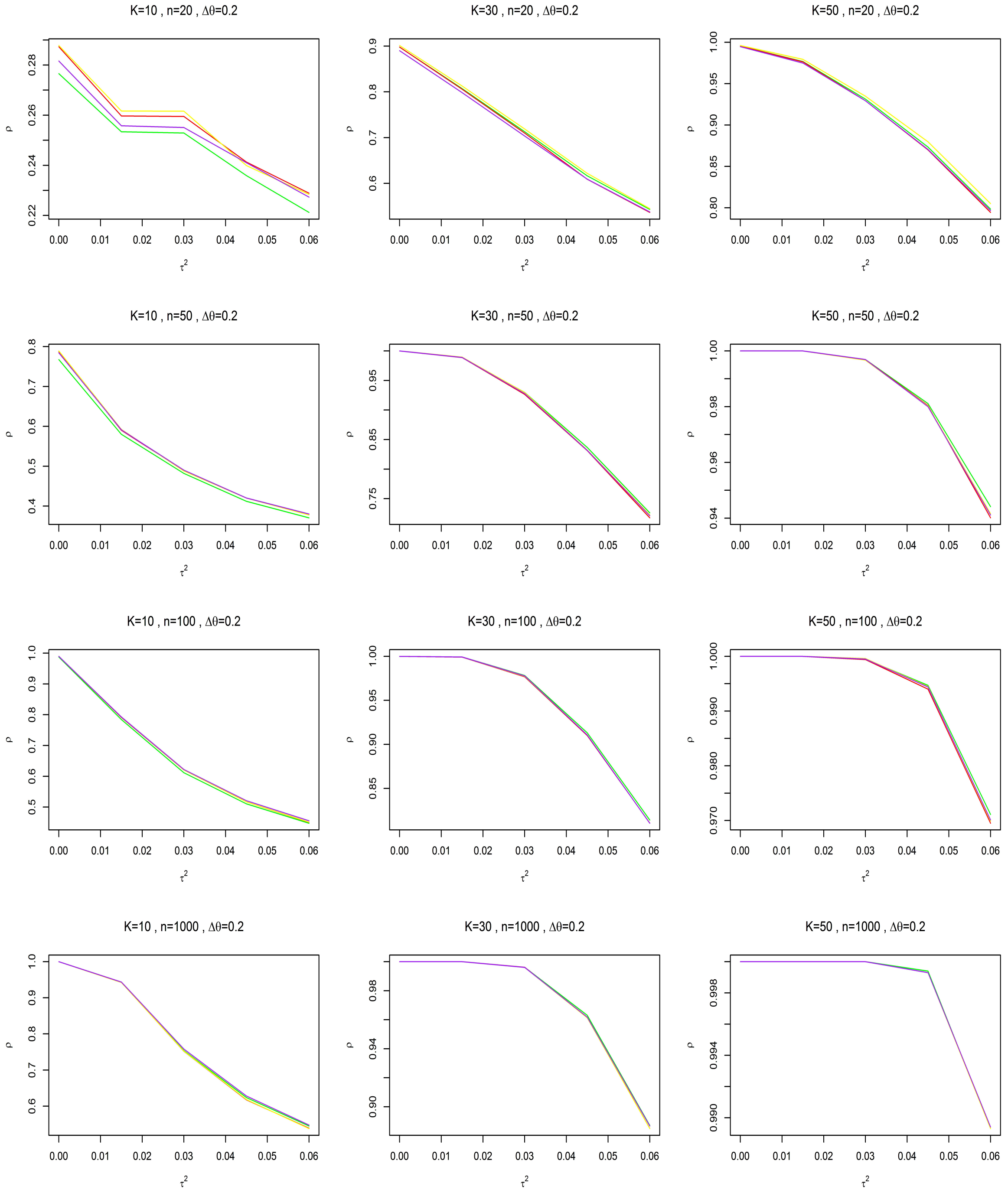


Figure 4.3: Power of Gombay test for REM with standard critical values against τ^2 based on DerSimonian and Laird (1986); Higgins et al. (2011); Paule and Mandel (1982) and REML estimators of τ^2 (GDL -red line, GH - green, GPM - yellow and GREML - purple line, respectively). K is the number of studies; n is the average studies size; ρ is the power and τ^2 is the between-study variance. $\Delta\theta$ is the amount of shift in value of effect parameter from θ_0.

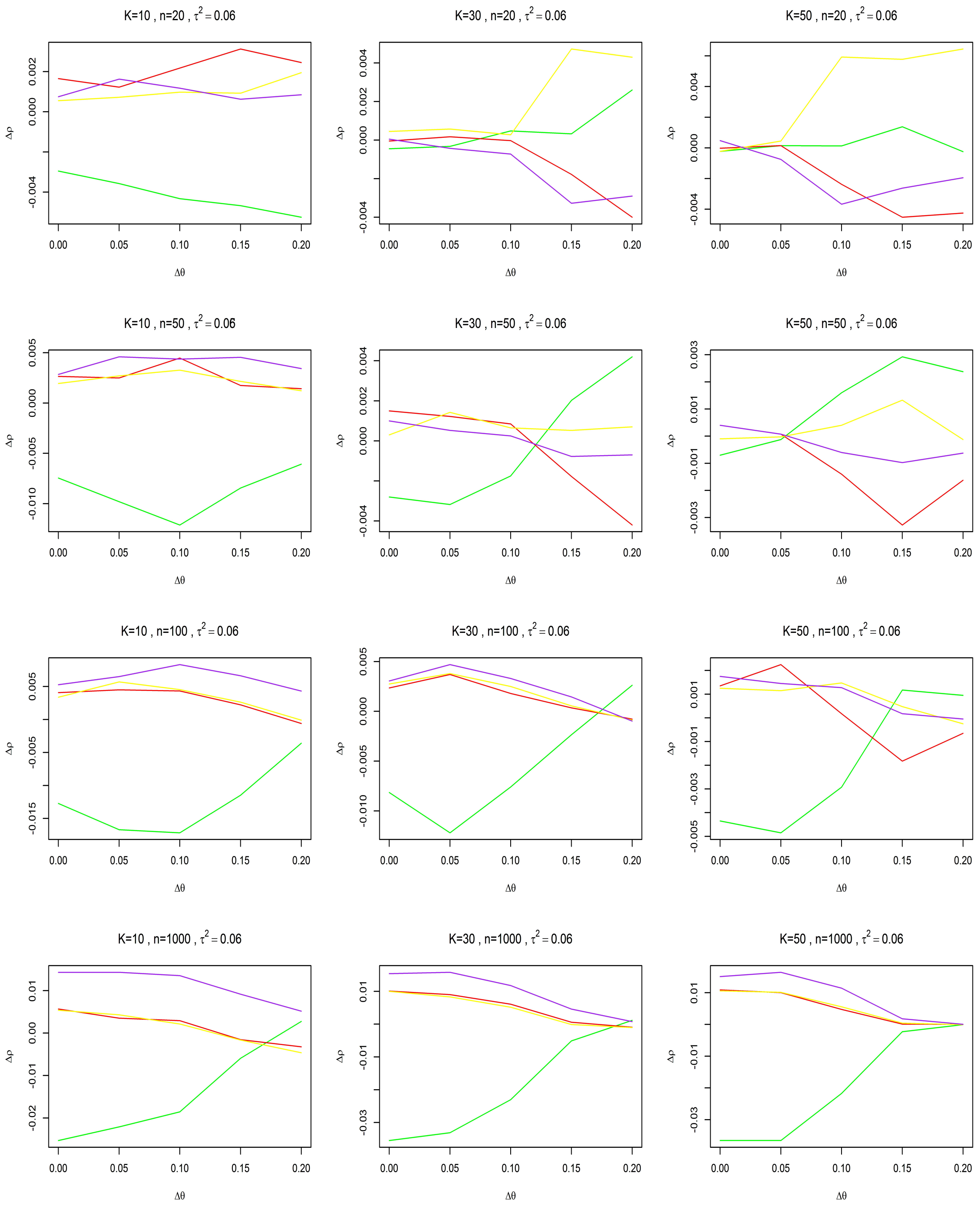


Figure 4.4: Deviations in power $\Delta\rho$ of Gombay test for REM with standard critical values at 5% level from the mean powers of the four tests based on DerSimonian and Laird (1986); Higgins et al. (2011), Paule and Mandel (1982) and REML estimators of τ^2 (GDL -red line, GH - green, GPM - yellow and GREML - purple line, respectively). K is the number of studies; n is the average studies size; $\Delta\rho$ is the deviation in power from the mean power of the tests based on 4 estimators of τ^2, $\Delta\theta$ is the amount of shift in value of the effect parameter from θ_0; τ^2 is the between-study variance.

of the variables used in the simulation were specifically chosen to cover most of the practical situations. See the results in Figures 4.1-4.4.

4.3.1 Type I Error of the Gombay test for REM

A key issue in testing of statistical hypothesis is the ability to achieve good power while maintaining the probability of Type I error, that is, the probability of false rejection. Figure 4.1 shows the overall Type I errors achieved by Gombay test for REM with standard critical values based on four estimators of τ^2; DerSimonian and Laird (1986); Higgins et al. (2011), Paule and Mandel (1982) and the REML (GDL, GH, GPM and GREML, respectively). When n=20, the values of Type I error rates achieved by the test based on all the four estimators are below the nominal level of 0.05. But as n increases to 50, GDL, GPM and GREML cross the nominal 5% level for larger values of τ^2. The achieved level of GH is still below the nominal level for all studied values of heterogeneity. When n=100, the Type I error rates achieved by the tests based on all the four estimators of τ^2 increase and cross the nominal level when $\tau^2 = 0.025$ for GDL, GPM and GREML and when $\tau^2 = 0.04$ for GH. For all values of n and τ^2, GDL, GPM and GREML produce higher Type I error rates compared to GH.

In general, the Gombay test for REM with standard critical values does not control the Type I error rate well. The Type I errors achieved by the Gombay method increase with increase in K, n and τ^2, and the tests do not control the Type I error rate. Besides, the levels achieved by the tests in FEM when $\tau^2 = 0$ are practically zero.

4.3.2 Statistical Power of Gombay test for REM

Figures 4.2-4.4 show the analysis of the power of Gombay test for REM based on DerSimonian and Laird (1986); Higgins et al. (2011), Paule and Mandel (1982) and the REML estimators of τ^2. As expected, the power increases with increase in the number of studies K, average study size n and the value of the population treatment effect θ. Figure 4.3 demonstrates that the power decreases with increase in heterogeneity τ^2. This should be expected as the increase in variability makes detection of a treatment effect more difficult. However, counter-intuitively the power increases with heterogeneity when n=20. The reason for this is the extreme conservativeness of the Gombay test for REM when n is relatively small, see Figure 4.1. Without the control of Type I error rate, a comparison of power is pointless. However, Figure 4.4 shows comparison of the power of the tests based on four different estimators of τ^2 when the value of $\tau^2 = 0.06$. The differences in the power between the four tests are very small. When $n = 20$ GREML is more powerful, followed by GDL, GMP and GH is the least powerful. To some extend this is also true for larger values of n, however as the value of θ increases, the power of GH increases and it eventually becomes more powerful compared to the other three tests.

4.4 Discussion

This Chapter has considered the use of asymptotic Gombay method for sequential meta-analysis that incorporates random effects and accounts for heterogeneity amount of the treatment effects. The Gombay test statistic for REM has been defined based on four different estimators of τ^2; DerSimonian and Laird (1986), Higgins et al. (2011), Paule and Mandel (1982) and REML methods. However, simulations of the Gombay test for REM with standard critical values obtained from asymptotic theory show that the test does not control the Type I error rate well. As was shown in the simulation results, the Type I error achieved by the test is close to zero when the value of τ^2 is small. In contrast, larger values of τ^2 lead to considerable inflation of the Type I error rate. The Type I error of the test also depends on the values of the average sample size n and the number of studies K in the analysis.

Without the control of type I error, the comparison of power of the tests based on different estimators of τ^2 is not valid, though the test based on REML estimator of τ^2 appears to result in the higher statistical power compared to the tests based on other three estimators considered.

In general, the use of Gombay method with the standard critical values obtained from asymptotic theory is disappointing. However the Gombay method has some important characteristics that can be improved upon to provide a better sequential approach for random-effects meta-analysis. In particular, the lack of control of the Type I error rate by the proposed test can be explained by the use of asymptotic approximations based on Wiener's process (see Gombay and Serbian (2005)) to obtain the critical values of the Gombay test.

Another problem that might have contributed to the lack of control of the Type I error rate by the proposed method is that the Gombay method assumed that the sample observations have the same distribution f. However, in REM the variances of estimated effects differ, and the sequence $\{T_k\}$ can by Wiener process only for very large within-studies sample sizes which make within-study variances σ_i^2 negligible. In the next Chapter, bootstrap critical values shall be determined for the use with the Gombay test. The results of this Chapter are published in International Journal of Mathematical, Computational, Statistical, Natural and Physical Engineering, see Dogo et al. (2015).

Chapter 5

Gombay test for REM with bootstrap critical values

In the previous Chapter, the Gombay method was introduced for sequential monitoring of temporal changes in effect sizes in random-effects meta-analysis. However, simulation has shown that Gombay test for REM with standard critical values does not control the Type I error rate well. The standard critical values of the Gombay method are determined based on asymptotic approximation of the distribution of the test statistic under the null hypothesis, see Gombay and Serbian (2005). Asymptotic theory often provides inaccurate approximation of finite sample distributions of test statistic (Horowitz, 1997). As follows from the simulations in Section 4.3, a poor approximation of the distribution of the Gombay test statistic for REM under the null hypothesis leads to a test that does not control the Type I error rate and has low statistical power. Bootstrap-based critical values are introduced in this Chapter for the use with the Gombay method. For simplicity, this test is referred to as retrospective Gombay sequential bootstrap test for REM.

The bootstrap is a method for estimating the distribution of an estimator or a test statistic by re-sampling the data (Horowitz, 2001). The data in bootstrap method is treated as if it were the population with the aim to evaluate the distribution of interest. The bootstrap is a computer-based method which substitutes considerable amounts of computation in place of theoretical analysis (Efron and Tibshirani, 1985). Bootstrap-based critical values produce spectacular reduction in the finite sample error compared to the asymptotic ones (Hall and Horowitz, 1996), and provide dramatic reductions in the difference between the true and nominal levels of a test (Horowitz, 1997). In Section 5.1, the retrospective Gombay sequential bootstrap test for REM is presented. Section 5.2 reports on simulations of retrospective Gombay sequential bootstrap test for REM. Sections 5.3 and 5.4 present the application and discussion of the use of retrospective Gombay sequential bootstrap test in random-effects model of meta-analysis, respectively.

5.1 The retrospective Gombay sequential bootstrap test for REM

Define the retrospective Gombay sequential bootstrap test statistic for REM for detection of a shift $\Delta\theta$ in the effect parameter from θ_0 by

$$G_K = ^{\max}_{1<k\leq K} \frac{1}{\sqrt{K}} \frac{\sum_{i=1}^{k} w_i^*(\hat{\tau}_K^2)(y_i-\theta_0)}{\sqrt{\sum_{i=1}^{k} w_i^*(\hat{\tau}_K^2)}}. \tag{5.1}$$

The statistics $\sum_{i=1}^{k} \hat{w}_i^*(y_i-\theta_0)$ and $\sum_{i=1}^{k} \hat{w}_i^*(\hat{\tau}_K^2)$ are the estimates of the efficient score vector and the value of the Fisher information at the k-th interim analysis calculated based on the best estimate of τ^2 from all available K studies, $\hat{\tau}_K^2$. Note that as the knowledge of $\hat{\tau}_K^2$ is required, this is not a true sequential test. This is rather a method allowing retrospective analysis of the sequential combined effects in random-effects meta-analysis.

5.1.1 Bootstrap procedure

Consider the following one- and two-sided retrospective tests for the existence of a shift from θ_0, say $\Delta\theta$. The tests are to be performed after combining K studies. Define T_k, for k=2, ..., K as

$$T_k = \frac{\sum_{i=1}^{k} \hat{w}_i^*(\hat{\tau}_K^2)(y_i-\theta_0)}{\sqrt{\sum_{i=1}^{k} \hat{w}_i^*(\hat{\tau}_K^2)}}, \tag{5.2}$$

Test: For k =2, 3, ..., K, reject H_0 if $T_k \geq KC(\alpha)$ (one-sided) or $|T_k| \geq KC_*(\alpha)$ (two-sided) and if no such k, $k \leq K$, exists do not reject H_0. The critical values $C(\alpha)$

and $C_*(\alpha)$ are to be calculated by bootstrap. Let

$$G^* =^{\max}_{2\leq k\leq K} \left\{ \frac{1}{\sqrt{K}} T_k \right\} \text{ and } G^{**} =^{\min}_{2\leq k\leq K} \left\{ \frac{1}{\sqrt{K}} T_k \right\}.$$

The calculation of the bootstrap critical values is based on the percentiles of the empirical distribution of G^* and G^{**} calculated from the set of bootstrap samples of the data. The step procedure for the calculations are as follows.

1. From the observed data, calculate the effect estimates y_i, the sample variances S_i^2, study sizes, n_i, and other sample statistics as required, for i=1, 2, ..., K. Calculate $\hat{\tau}_K^2$ using one of the methods for estimating τ^2 by DerSimonian and Laird (1986), Higgins et al. (2011), Paule and Mandel (1982) or REML.

2. Use the values of $\hat{\tau}_K^2$, the null value of the effect parameter, θ_0 and other sample statistics to draw B independent bootstrap samples of the effect estimates from an appropriate distribution, i.e. the distribution of the sampled data from studies. Calculate or generate from an appropriate distribution B bootstrap estimates of the within-study variances, $S_{b_i}^2$, for i=1, 2, ..., K. A standard choice for constructing bootstrap test is to use B=1000.

3. Use the bootstrap values $\{(y_{b_i},\ S_{b_i}^2)$, i=1, 2, ..., K$\}$ to calculate the estimate of τ^2, $\hat{\tau}_b^2$ for each sample b=1, 2, ..., B, and the corresponding estimated weights in random-effects model as $w_{b_i} = (\hat{\tau}_b^2 + S_{b_i}^2)^{-1}$.

4. For each bootstrap sample b =1, 2, ..., B, calculate the sequential statistics

$$T_{bk} = \sum_{i=1}^{k} w_{bi}^*(y_{bi} - \theta_0) / \sqrt{\sum_{i=1}^{k} w_{bi}^*}. \tag{5.3}$$

Find G_b^* and G_b^{**} statistics as follows.

$$G_b^* =_{2\leq k\leq K}^{max} \left\{K^{-\frac{1}{2}}T_{bk}\right\};\ G_b^{**} =_{2\leq k\leq K}^{min} \left\{K^{-\frac{1}{2}}T_{bk}\right\} \tag{5.4}$$

5. Order the bootstrap replicates of G_b^* and G_b^{**}, as $G_1^* \leq G_2^* \leq G_3^* \leq ... \leq G_B^*$ and $G_1^{**} \leq G_2^{**} \leq G_3^{**} \leq ... \leq G_B^{**}$. For a one-sided test, the upper critical values are given by $[B \times (1-\alpha)+1]^{th}$ element in the sequence of $\{G_b^*\}$, while the lower critical values are calculated by $[B \times \alpha]^{th}$ element in the sequence of $\{G^{**}\}$. Use $\alpha/2$ instead of α for the two-sided test.

Step 2 of the above bootstrap procedure is very effect measure specific. Below are presented the details for several popular effect measures available in the R program provided in the Appendix.

5.1.2 Sample mean

When the effect of interest y_i is the sample mean of the n_i normally distributed observations, and its variances $S_i^2 = s_i^2/n_i$ for the sample variance s_i^2, generate B bootstrap effects $y_{bi} \sim N(\theta_0, \hat{\tau}^2 + S_i^2)$ and B bootstrap estimates of the within-study variances, $S_{ib}^2 \sim S_i^2\chi_{n_i-1}^2$, for $i = 1,\ 2,\ ...,\ K$.

5.1.2.1 Mean difference

When the effect of interest y_i is the difference of the treatment (T) and control (C) sample means of normally distributed observations, denote sample variances in the two arms by s_{iT}^2 and s_{iC}^2, with the sample sizes n_{iT} and n_{iC}, respectively. The variance of the mean difference is $S_i^2 = s_{iT}^2/n_{iT} + s_{iC}^2/n_{iC}$. Generate B bootstrap effects $y_{bi} \sim N(\theta_0, \hat{\tau}^2 + S_i^2)$ and B pairs of the bootstrap within-arms sample variances, $s_{biT}^2 \sim s_{iT}^2 \chi^2_{n_{iT}-1}/(n_{iT}-1)$ and $s_{biC}^2 \sim s_{iC}^2 \chi^2_{n_{iC}-1}/(n_{iC}-1)$, for $i = 1, ..., K$. Then calculate the within-studies variances, $S_{bi}^2 = s_{biT}^2/n_{iT} + s_{biC}^2/n_{iC}$.

5.1.2.2 Standardised mean difference

The standardised mean difference (SMD) $\delta = (\mu_{Ti} - \mu_{Ci})/\sigma$ is estimated by $y_i = J(N_i)(\bar{X}_{Ti} - \bar{X}_{Ci})/s_{pi}$ for the pooled sample variances $s_{pi}^2 = [(n_{iT}-1)s_{iT}^2 + (n_{iC} - 1)s_{iC}^2]/(N_i - 2)$, where $N_i = n_{Ci} + n_{Ti}$ and $J = \Gamma[(N_i-2)/2]/(\sqrt{(N_i-2)/2}\Gamma[(N_i-3)/2])$ is a constant depending only on the total sample size N_i. The variance is estimated by (see (Hedges and Olkin (1985), p. 104-5))

$$S_i^2 = \frac{(N_i-2)N_i J_i^2}{(N_i-4)n_{Ci}n_{Ti}} + \left(\frac{(N_i-2)J_i^2}{N_i-4} - 1\right) y_i^2 := A_i + B_i y_i^2, \tag{5.5}$$

where A and B depend only on sample sizes. $\sqrt{n_{Ti}n_{Ci}/N_i}y_i$ has non-central t-distribution with $N_i - 2$ degrees of freedom, and the non-centrality parameter $\sqrt{n_{Ti}n_{Ci}/N_i}\delta$, denoted by $t(N_i-2, \sqrt{n_{Ti}n_{Ci}/N_i}\delta)$. Generate B bootstrap effects y_{bi} from $[J(N_i)N_i/\sqrt{n_{Ti}n_{Ci}}] \times t(N_i - 2, \sqrt{n_{Ti}n_{Ci}/N_i}\delta)$ distribution and calculate their variances S_{bi}^2 from equation (5.5).

5.1.3 Binomial effect measures

Denote the numbers of events in the control and treatment arms of the studies by X_{Ci} and X_{Ti}, respectively. Let $a = 0$. When $X_{Ci} = 0$ or $X_{Ci} = n_{Ci}$, take $a = 1/2$. Estimate probabilities $p_{Ci} = (X_{Ci} - a)/(n_{Ci} + 2a)$. When the effect of interest y_i is the log odds ratio or the log relative risk, discard the studies with $X_{Ci} + X_{Ti} = 0$ or $n_{Ci} + n_{Ti}$ and adjust the total number of studies K accordingly.

5.1.3.1 Log odds ratio

When the effect of interest y_i is the log odds ratio, generate B vectors of length K containing within-study log odds ratios $\theta_{bi} \sim N(\theta_0, \hat{\tau}^2)$. Given the values of p_{Ci} and θ_{bi}, the logits in the treatment groups are $\text{logit}(p_{T_{bi}}) + \theta_{bi}$. Hence calculate the probabilities p_{Tbi} and simulate the numbers of the study outcomes X_{Tbi} and X_{Cbi} from the binomial distributions $Binom(n_{Ti}, p_{Tbi})$ and $Binom(n_{Ci}, p_{Ci})$, respectively. Following Gart et al. (1985) to obtain unbiased estimators of the log odds ratios and their variances, calculate the log odds ratios as $y_{bi} = \log[(X_{Tbi} + 1/2)/(n_{Ti} - X_{Tbi} + 1/2)] - \log[(X_{Cbi} + 1/2)/(n_{Ci} - X_{Cbi} + 1/2)]$ and the variances are $S^2_{bi} = (X_{Tbi} + 1/2)^{-1} + (n_{Ti} - X_{Tbi} + 1/2)^{-1} + (X_{Cbi} + 1/2)^{-1} + (n_{Cbi} - X_{Cbi} + 1/2)^{-1}$.

5.1.3.2 Log relative risk

In case of log relative risks y_i, generate B vectors of length K containing within-study mean log relative risks θ_{bi} from $N(\theta_0, \hat{\tau}^2)$ distribution with the i-th distribution truncated on the left at $-\log p_{Ci}$, $i = 1$, ..., K. The use of truncated normal distributions to restrict the range of the possible values of log relative risks is required to guarantee the treatment probabilities below 1. Given the values of p_{Ci} and θ_{bi}, calculate the probabilities in the treatment groups $p_{Tbi} = p_{Ci} exp(\theta_{bi})$. Generate the numbers of the study outcomes X_{Tbi} and X_{Cbi} from the binomial distributions $Binom(n_{Ti}, p_{Tbi})$ and $Binom(n_{Ci}, p_{Ci})$, respectively. Following Pettigrew et al. (1986) to obtain the unbiased estimators of the log relative risk and their variances, calculate the log relative risk as $\log((X_{Tbi} + 1/2)/(n_{Ti} + 1/2)) + \log((X_{Cbi} + 1/2)/(n_{Ci} + 1/2))$ and the variances as $S_{bi}^2 = (X_{Tbi} + 1/2)^{-1} - (n_{Ti} + 1/2)^{-1} + (X_{Cbi} + 1/2)^{-1} - (n_{Ci} + 1/2)^{-1}$.

5.1.3.3 Risk difference

In case of risk difference y_i, generate B vectors of length K containing within-study mean risk differences θ_{bi} from $N(\theta_0, \hat{\tau}^2)$ distributions with the ith distribution truncated to the interval

[-$p_{Ci}, 1 - p_{Ci}$], i = 1, ..., K. The use of truncated normal distributions to restrict the range of possible values of risk differences is required to guarantee the treatment probabilities below 1. Given the values of p_{Ci} and θ_{bi}, calculate the probabilities in the treatment groups $p_{Tbi} = p_{Ci} + \theta_{bi}$. Generate the numbers of the study outcomes X_{Tbi} and X_{Cbi} from the binomial distributions $Binom(n_{Ti}, p_{Tbi})$ and $Binom(n_{Ci}, p_{Ci})$, respectively. Calculate the risk differences as $y_{bi} = X_{Tbi}/n_{Ti} - X_{Cbi}/n_{Ci}$ and the variance as $S_{bi}^2 = (X_{Tbi} + a)(n_{Ti} - X_{Tbi} + a)/(n_{Ti} - 2a)^3 + (X_{Cbi} + a)(n_{Ci} - X_{Cbi} + a)/(n_{Ci} - 2a)^3$. Use $a = 0$ unless $X_{Tbi} = 0$ or $X_{Tbi} = n_{Ti}$ or $X_{Cbi} = 0$ or $X_{Cbi} = n_{Ci}$, in which case use $a = 1/2$.

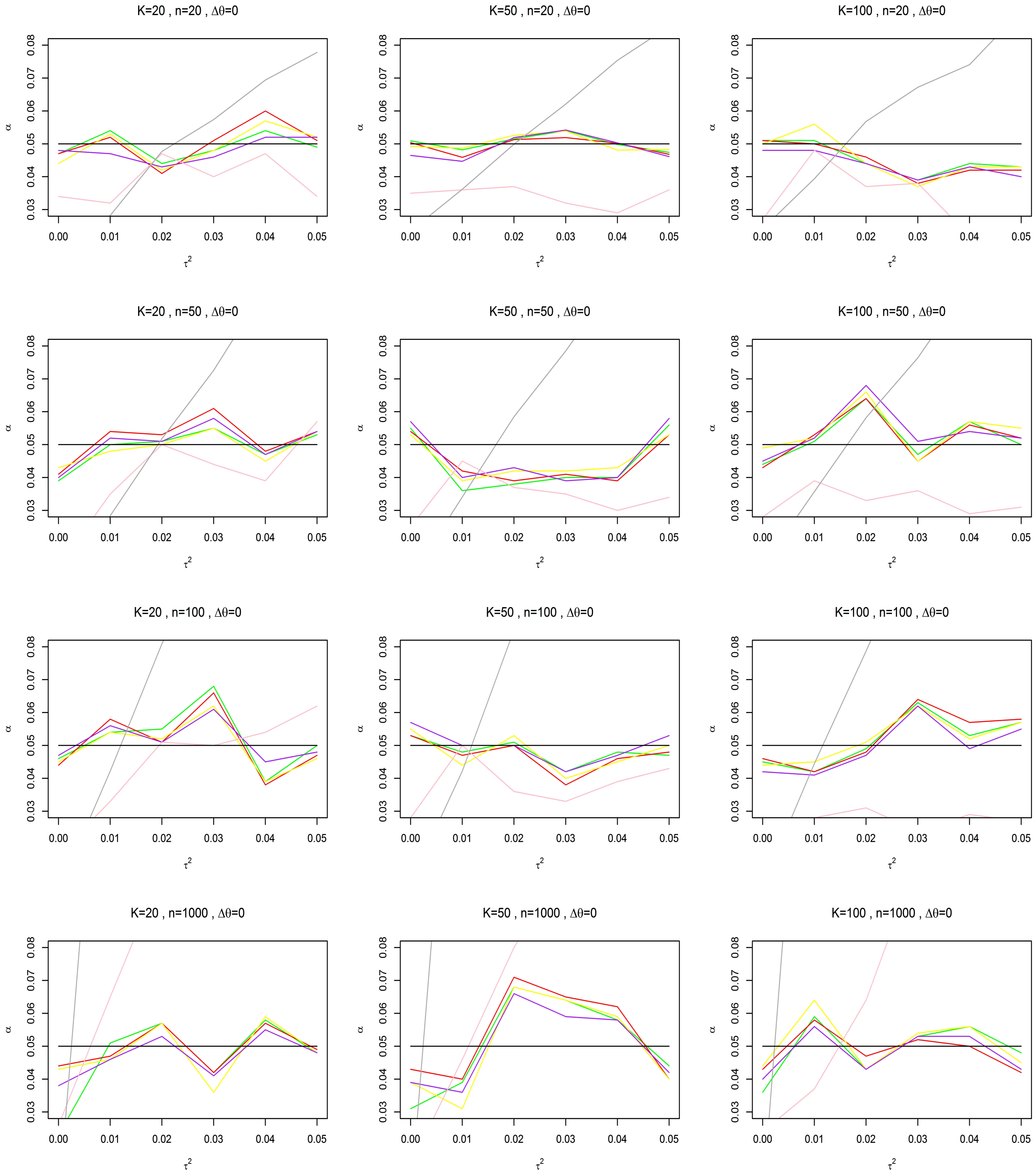


Figure 5.1: Type I errors achieved by retrospective Gombay sequential bootstrap test for REM based on DerSimonian and Laird (1986), Higgins et al. (2011), Paule and Mandel (1982) and REML estimators of τ^2 (GDL -red line, GH - green, GPM - yellow, GREML - purple line, penalized Z-test-darkgrey, and SMA based on Pocock's boundaries-pink, respectively) . K is the number of studies; n is the average sample size; $\Delta\theta$ is the shift in effect parameter from $\theta_0 = 0$, τ^2 is the between-study variance. The black straight line represent the nominal level of 5% for the test.

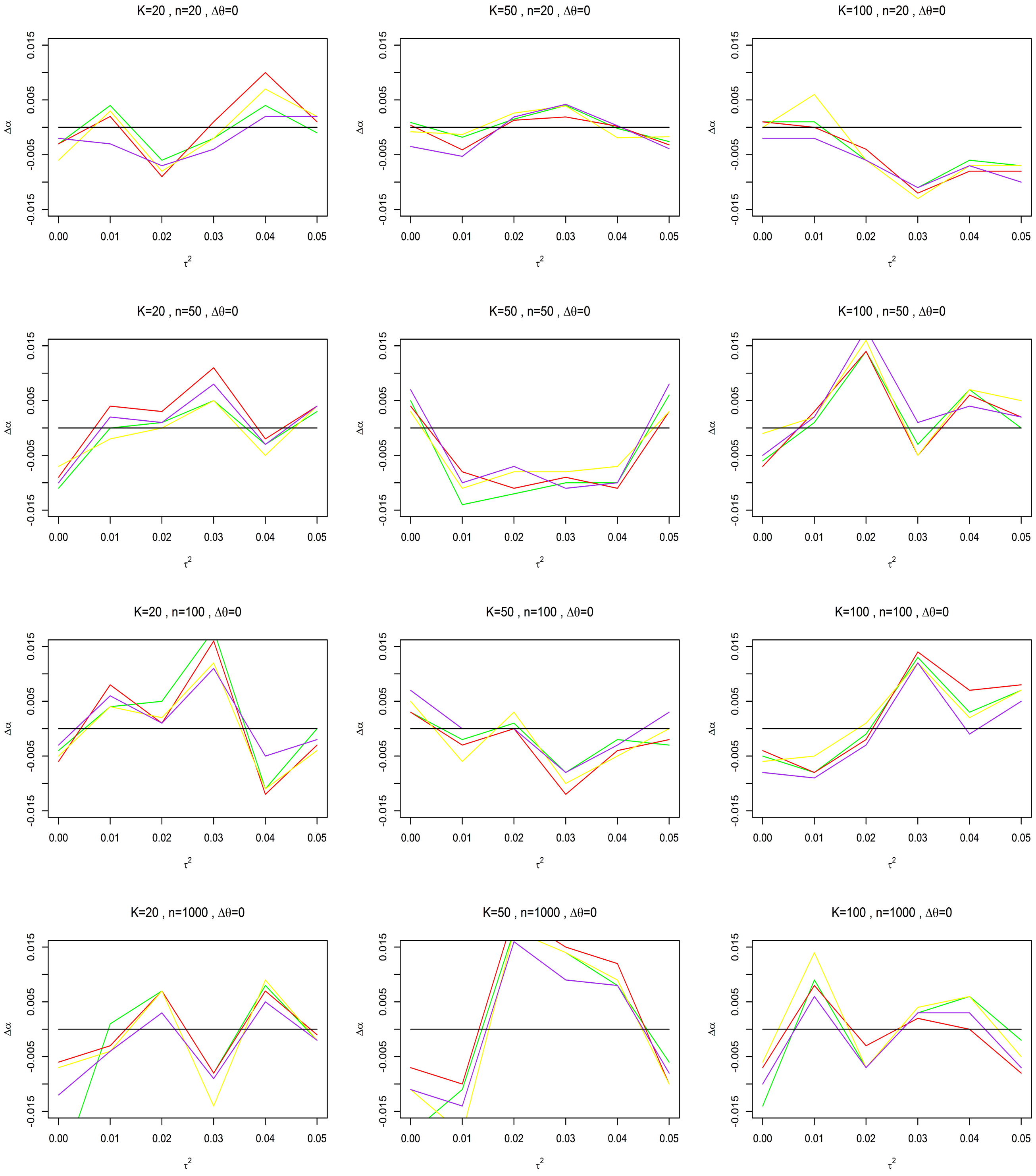


Figure 5.2: Deviation of Type I error $\Delta\alpha$ achieved by retrospective Gombay sequential bootstrap tests for REM from the nominal level based on DerSimonian and Laird (1986), Higgins et al. (2011), Paule and Mandel (1982) and the REML estimators of τ^2 (GDL -red line, GH - green, GPM - yellow and GREML - purple line, respectively). K is the number of studies; n is the average sample size; $\Delta\theta$ is the shift in effect parameter from $\theta_0 = 0$, τ^2 is the between-study variance and nominal level of the tests is $\alpha = 0.05$. The black straight line corresponds to point where the difference is zero.

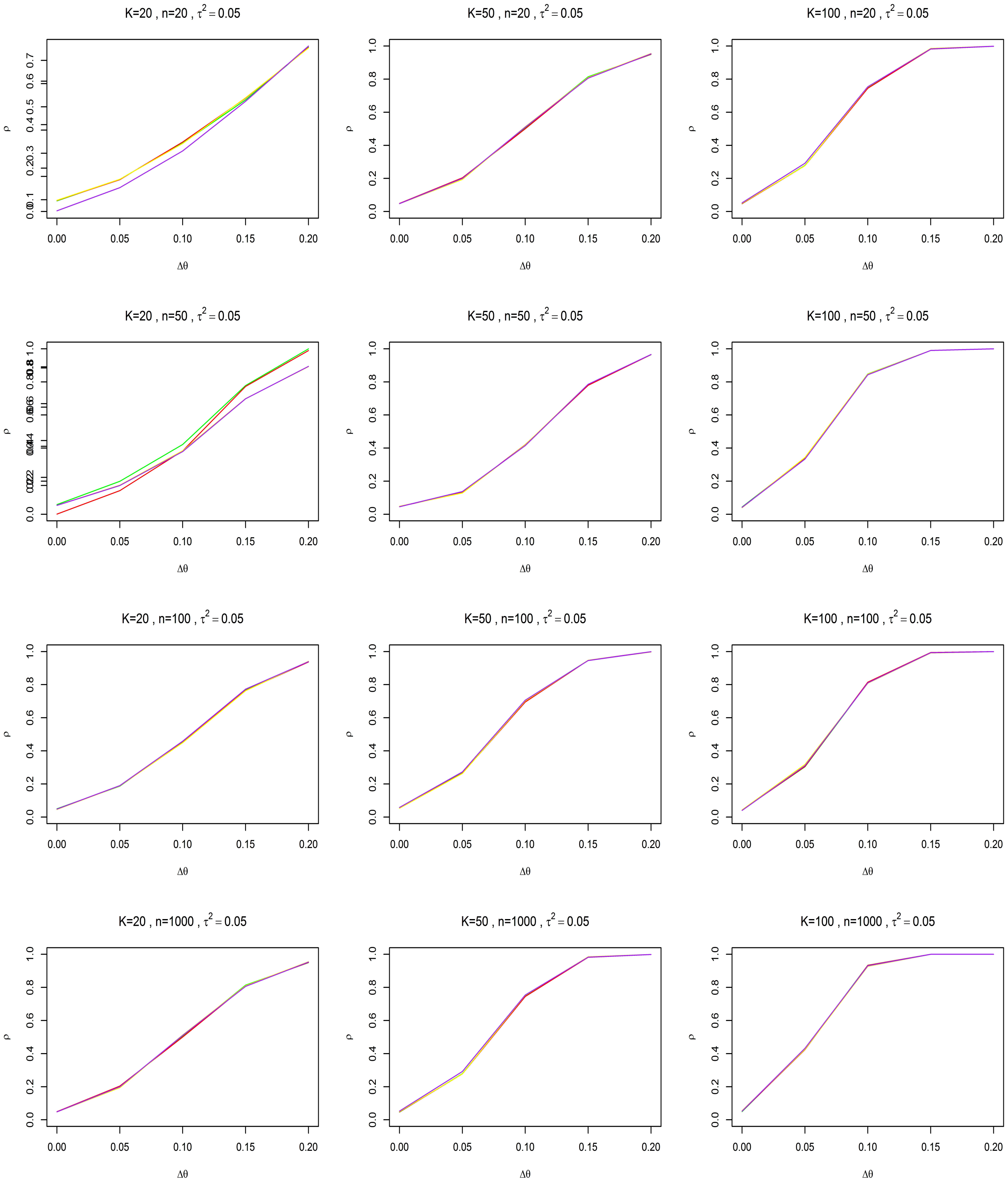


Figure 5.3: The power of retrospective Gombay sequential bootstrap test for REM based on DerSimonian and Laird (1986), Higgins et al. (2011), Paule and Mandel (1982) and REML estimators of τ^2 (GDL -red line, GH - green, GPM - yellow and GREML - purple line, respectively) against $\Delta\theta$. K is the number of studies; n is the average sample size; ρ is the power while $\Delta\theta$ is the change in effect parameter from $\theta_0 = 0$, τ^2 is the between-study variance.

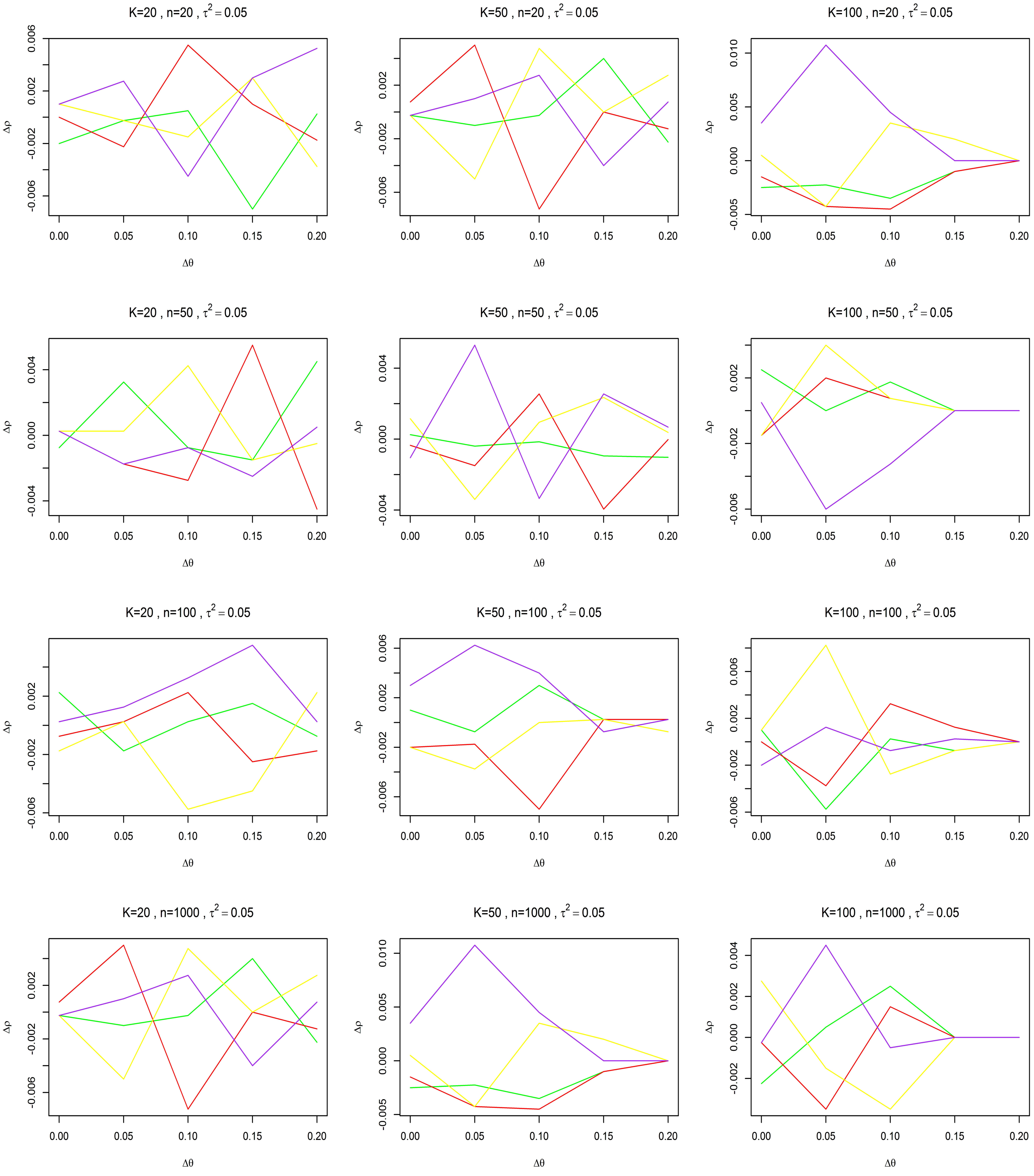


Figure 5.4: Deviations of power of retrospective Gombay sequential bootstrap test for REM from the mean power of the tests based on DerSimonian and Laird (1986), Higgins et al. (2011), Paule and Mandel (1982) and REML estimators of τ^2 (GDL -red line, GH - green, GPM - yellow and GREML - purple line, respectively) when $\tau^2 = 0.05$. K is the number of studies; n is the average sample size; $\Delta\rho$ is the deviations in power from the average power of the four tests while $\Delta\theta$ is the change in effect parameter from $\theta_0 = 0$, τ^2 is the between-study variance.

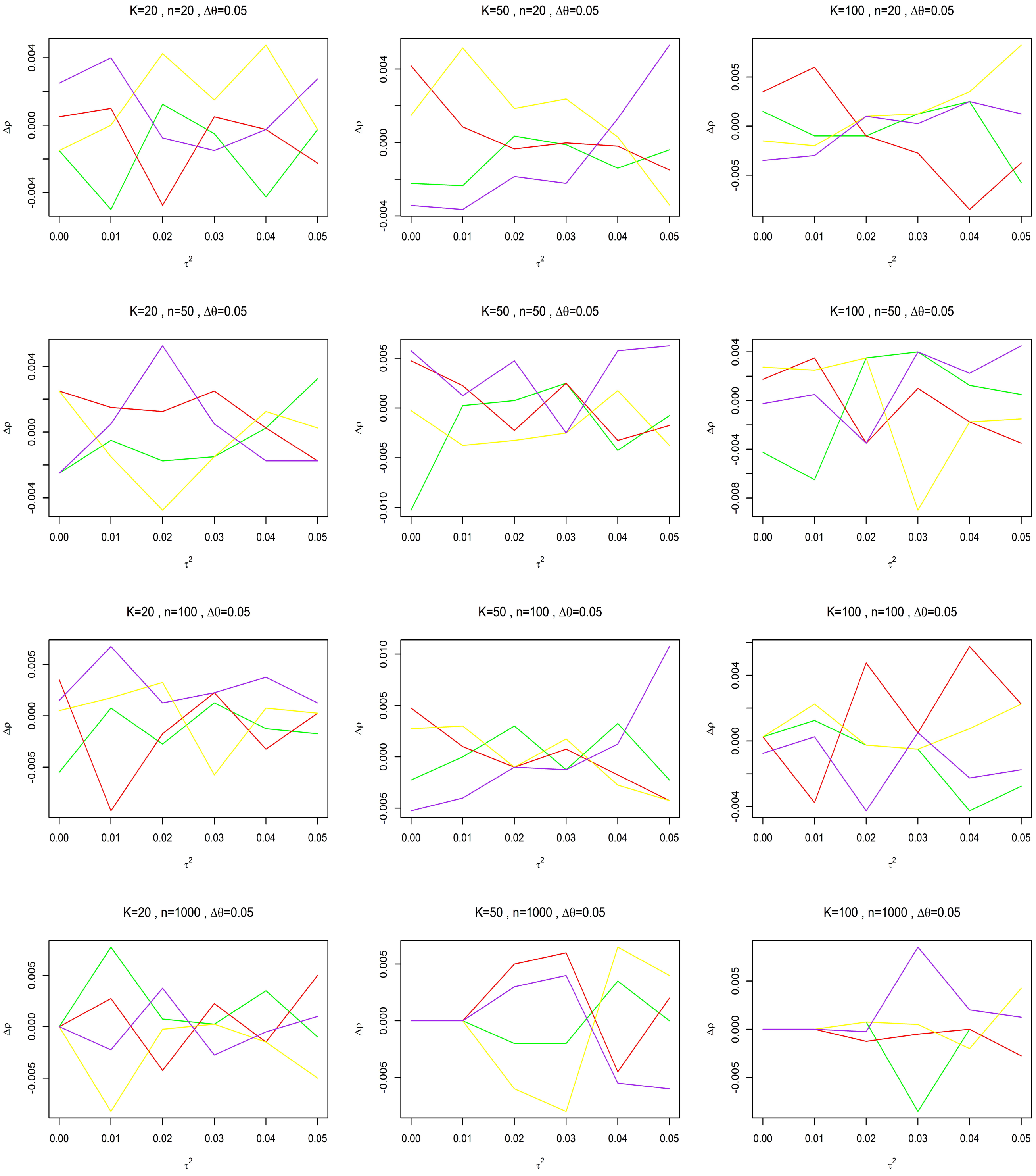


Figure 5.5: Deviations of power of retrospective Gombay sequential bootstrap test for REM from the mean power of the tests based on DerSimonian and Laird (1986), Higgins et al. (2011), Paule and Mandel (1982) and REML estimators of τ^2 (GDL -red line, GH - green, GPM - yellow and GREML - purple line, respectively) when $\Delta\theta = 0.05$. K is the number of studies; n is the average sample size; $\Delta\rho$ is the deviations in power from the average power of the four tests while $\Delta\theta$ is the change in effect parameter from $\theta_0 = 0$, τ^2 is the between-study variance.

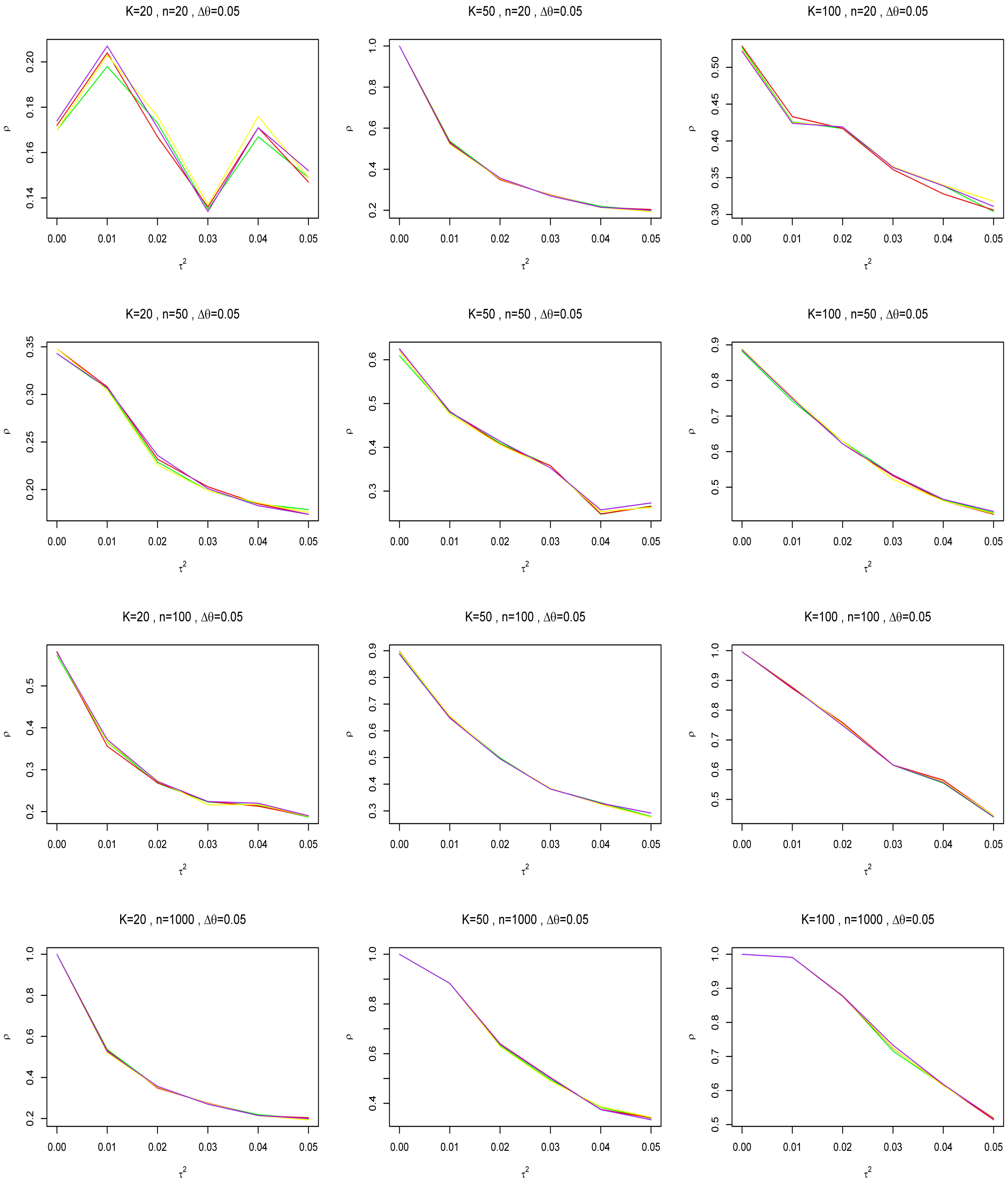


Figure 5.6: The power of retrospective Gombay sequential bootstrap test for REM based on DerSimonian and Laird (1986), Higgins et al. (2011), Paule and Mandel (1982) and REML estimators of τ^2 (GDL -red line, GH - green, GPM - yellow and GREML - purple line, respectively) when $\Delta\theta = 0.05$ against τ^2. K is the number of studies; n is the average sample size; ρ is the power while $\Delta\theta$ is the change in effect parameter from $\theta_0 = 0$, τ^2 is the between-study variance.

5.2 Simulation Study

To evaluate the properties of the retrospective Gombay sequential bootstrap test presented in Section 5.1, a simulation study was conducted. The observed estimates of the treatment effect were generated using the normal distribution, $y_i \sim N(\Delta\theta, \tau^2 + \sigma_i^2)$, where $\Delta\theta$ is the difference in the null value of effect parameter θ_0 and the alternative $\theta_0 + h$. The studies sizes were generated using the normal distribution, $n_i \sim N\left(n, \frac{n}{4}\right)$ rounded to the nearest integer and truncated on the left at 3, n is the average sample size of the studies. Estimates of sample variances, $\hat{\sigma}_i^2$ were generated using scaled Chi-square distributions, $\hat{\sigma}_i^2 \sim \frac{\sigma_i^2}{(n_i-1)}\chi^2_{n_i-1}$. This choice ensures that $\mathrm{E}[\hat{\sigma}_i^2] = \sigma_i^2$. Estimated variances of estimated treatment effects y_i are $S_i^2 = \hat{\sigma}_i^2/n_i$. The data for each simulated meta-analysis consisted of a total of K estimates of the observed treatment effects, their estimated variances, and corresponding sample sizes $\{(y_i, S_i^2, n_i), \quad i = 1, \cdots, K\}$. For each dataset four bootstrap-based tests were calculated using different estimators of τ^2: DerSimonian and Laird (1986), Higgins et al. (2011), Paule and Mandel (1982) and REML (GH, GDL, GPM and GREML, respectively), the penalized Z-test by Lan et al. (2003) with $\lambda = 2$, and SMA based on Lan-DeMets alpha-spending function (Lan and DeMets, 1983) and Pocock's boundaries as implemented in program *ldbands* from the R package *Hmisc* (Casper and Perez, 2006). Following Wetterslev et al. (2008), the OIS for SMA was inflated by an adjustment factor $(1 - I^2)^{-1}$ for the I^2 inconsistency index $I^2 = (Q - (K-1))/Q$ (this method is referred to as SMA in the rest of the paper). We used one-sided tests and the significance level was fixed at $\alpha = 0.05$. The null value of the effect parameter was taken as $\theta_0 = 0$ and the calculation of each bootstrap critical value was based on $B = 1000$ bootstrap replications. We generated 1000 datasets for each of the 270 combinations of the following variables chosen to represent a realistic range of the parameters values:

$\sigma^2 = 1$,

$\Delta\theta = (0.00, 0.05, 0.10, 0.15, 0.20)$,

$n = (20, 50, 100, 1000)$,

$K = (20, 50, 100)$ and

$\tau^2 = (0.00, 0.01, 0.02, 0.03, 0.04, 0.05)$.

For each scenario the number of times the test rejects the null hypothesis was recorded. The results are presented in Figures 5.1-5.6.

Figures 5.1 compare the overall Type I error rates achieved by retrospective Gombay sequential bootstrap test for REM based on Higgins et al. (2011), DerSimonian and Laird (1986), Paule and Mandel (1982) and REML estimators of τ^2 (GH, GDL, GPM and GREML), the penalised Z-test and SMA. Type I error rates in bootstrap-based tests with all the four estimators of τ^2 are relatively stable and close to the nominal level. When $K = 20$, the values of Type I error rates achieved by GH and GDL are somewhat higher compared to GPM and GREML, but as K increases to 50 and 100 there is very little difference between the four tests, as is clearer from Figure 5.2. Overall, even though there is no clear-cut winners, it appears that the GPM performs slightly better for smaller studies, and the GREML for large studies. In contrast, the Type I error rates for the penalised Z-test and the SMA are unsatisfactory. They are far from nominal value of 5% and increase with increasing values of K, n and τ^2. Interestingly, the SMA Type I error rate is mostly below the nominal level and seems to be unstable when $n \leq 100$ and $K \geq 50$, but it explodes with increasing τ^2 and $n = 1000$.

Figures 5.3- 5.6 show the analysis of power of the retrospective Gombay sequential bootstrap test for REM based on DerSimonian and Laird (1986), Higgins et al. (2011), Paule and Mandel (1982) and REML estimators of τ^2. In Figure 5.3, as to be expected, the power of the test increases with increase in number of studies K, average studies size n and the population treatment effect θ. In contrast, Figure 5.6 shows that power decreases with increase in heterogeneity variance τ^2. This also should be expected because increase in variability is known to result to wider confidence interval thereby making it difficult to detect the presence of an effect especially when it is small. Figure 5.4 compares power between the four tests when $\tau^2 = 0.03$. For $K = 20$ the results show that GH and GDL are more powerful compared to GPM and GREML; when $K = 50$ no clear difference in power is observed between the four tests, and as K increases to 100, GREML becomes more powerful compared to other three tests. In Figure 5.5 when $\tau^2 = 0.05$, no clear difference is observed anywhere in terms of the power of the four tests. The power of the retrospective Gombay sequential bootstrap tests for REM based on all the four different estimators is approximately the same.

Bias of the Type I error rate achieved in our simulations are as follows. Let $T_K = T_K(X_1, ..., X_K)$ be a statistic for testing H_0 and $C_K(\alpha)$ be the corresponding bootstrap critical value. If T_K is pivotal statistic, i.e. a statistic whose distribution does not depend on unknown parameter(s), Hall and Horowitz (1996), and Horowitz and Savin (2000) showed that

$$\mathrm{P}\left\{|T_K| > C_K(\alpha)\right\} = \alpha + O(K^{-2}), \tag{5.6}$$

where α is the nominal level. Since our retrospective Gombay bootstrap test statistic is asymptotically pivotal and K=(20,50,100), the maximum estimate of the bias in the Type I error rate achieved is 0.0025 which is negligible. Similarity between the results of the tests based on the different estimators of τ^2 can be explained by the fact that all the estimators were of a similar quality in regards to bias and precision when estimating τ^2.

To summarise on the basis of the simulations, the use of Gombay method with bootstrap critical values provides a remarkable reduction in the difference between the true and nominal levels of the test in comparable to the Gombay test for REM with critical values derived from asymptotic method. The test controls the Type I error rate well irrespective of the number of studies, studies sizes, the amount of heterogeneity in the treatment effects or the method of estimating τ^2, and also has high statistical power. Therefore, this research concentrates on the application of the retrospective Gombay sequential bootstrap test for REM in the following two examples of medical meta-analysis.

5.3 Examples

To demonstrate the application of the retrospective sequential bootstrap tests, this section consider two examples of medical meta-analyses. The results of the bootstrap tests are compared with the results obtained from CMA, CUSUM and SMA. The data for each meta-analysis were sorted chronologically according to year of publication, from the earliest to the latest. Where the year of publication of two or more studies coincide, the order was selected randomly. Cumulative meta-analysis were conducted using R package metafor (Viechtbauer, 2010). SMA was based on Lan-DeMets alpha-spending function (Lan and DeMets, 1983) and Pocock's boundaries as implemented in the R package ldbounds (Casper and Perez, 2006). CUSUM charts were obtained from the R package qcc (Scrucca, 2004).

5.3.1 Magnesium for myocardial infarction

The first application is based on the systematic review conducted by Li et al. (2007) to examine the effectiveness of the use of intravenous magnesium for the treatment of acute myocardial infarction. For simplicity, the data is referred to as the magnesium data. The data consist of 23 trials published from 1984 to 2004. The outcome of interest is mortality from acute myocardial infarction and the treatment effects are recorded as log-odds ratios. A correction factor 0.5 was added to each entry in the data and the log-odds ratios and its variances were calculated by

$$\hat{\varphi}_i = \log\left[\tfrac{(x_T+0.5)(n_C-x_C)}{(x_C+0.5)(n_T-x_T)}\right];\ \sigma_i^2(\hat{\varphi}) = \tfrac{1}{x_T+0.5} + \tfrac{1}{n_T-x_T} + \tfrac{1}{x_C+0.5} + \tfrac{1}{n_C-x_C}. \tag{5.7}$$

A negative value of φ_i indicates that mortality has been reduced and therefore favours the use of intravenous magnesium. A standard random effects meta-analysis of the data indicates a significant benefit in the use of magnesium with combined effect -0.2644 (p-value=0.0015), $\hat{\tau}^2_{DL} = 0.037$ and the value of Q-statistic is equal to 56.1405 with p-value< 0.0001. The data and the results of the analysis are presented in Tables 8.1 and 8.2 in the Appendix.

Let φ be the true value of the estimates of the log-odds ratio, $\hat{\varphi}_i$. For testing the effectiveness of the new intervention, consider the null hypothesis of no effect of intravenous magnesium, H_0: $\varphi = 0$. The CMA based on random-effects model and DerSimonian and Laird (1986) estimator of τ^2 at the target value of 0 first indicates significant effect with value -1.01 (p-value=0.016) at trial 3. However this result may be spurious due to the inflated Type I error rate in CMA. The CUSUM, SMA and the penalized Z-test with the same target value indicate a significant effect at trial 7. When the bootstrap based tests are used with the same target value of 0. GH and GDL reject H_0 at trials 5 with statistics values -0.4990 and -0.4984, respectively; while GPM and GREML reject H_0 at trials 6 with statistics values -0.5156 and -0.4892, respectively, see Figure 5.8. Hence for this data the bootstrap-based tests are more powerful in comparison to the CUSUM, SMA and the penalised Z-test.

Having established that intravenous magnesium is a significantly effective for acute myocardial infarction, it is important to monitor for any possible trend in the effect over time. In Figure 5.7, the beginning of an upward trend in the effectiveness of the use of intravenous magnesium for the treatment of acute myocardial infarction, so the CMA at this stage, -0.934 (cumulative log-odds ratio at trial 7) is set as the new target value. Figure 5.7 shows the analysis of the magnesium data using CMA based on random-effects model and SMA based on Pocock's boundaries, the penalised

Z-test and CUSUM. The horizontal line across the CMA corresponds to the combined log-odds ratio of -0.934 at stage 7. The CMA plot on Figure 5.7 exhibits a gradual increase in effect (corresponding to reduction in survival benefit), and the deviation from the horizontal line at -0.934 becomes significant at trial 10. The CUSUM chart also indicates the significant change at trial 10. The penalised Z-test (Hu et al., 2007, Lan et al., 2003) with the same target value crosses the upper boundary at trial 14 and SMA at trial 15. In Figure 5.9, GH and GDL methods indicate significant change at trial 15, while GPM and GREML indicate significant change at trials 20 and 22, respectively. As to be expected, the CMA and the CUSUM are liberal since they are based on fixed effect boundaries. The performance of the bootstrap based tests is consistent with the conclusion in the simulation study that GH and GDL are more liberal tests compared to GPM and GREML when the number of studies is not large.

5.3.2 Nicotine replacement therapy for smoking cessation

The second example is based on the systematic review by Stead et al. (2008) on testing the effectiveness of nicotine replacement therapy (NRT) for smoking cessation. Kulinskaya and Koricheva (2010) have reproduced and analysed the data using QC charts, and detected temporal changes in the effect of nicotine on smoking cessation. It will be interesting to see if such changes can be detected by retrospective Gombay sequential bootstrap test for REM. The data consist of 53 trials published from 1979 to 2005. The outcome of interest is the effect of nicotine containing chewing gum compared to control in aiding smoking cessation. The effect measure used is the log-relative risk estimated by

$$\hat{\phi}_i = \log\left[\frac{x_T.n_C}{x_C.n_T}\right] \text{ with variance estimated by } \sigma_i^2(\hat{\phi}) = \frac{n_T - x_T}{x_T.n_T} + \frac{n_C - x_C}{x_C.n_C}. \quad (5.8)$$

A positive value of $\hat{\phi}_i$ means that NRT is effective for smoking cessation. A random effects meta-analysis based on DerSimonian and Laird (1986) estimator of τ^2 indicates a significantly different from zero; log relative risk of 0.36 (RR=1.43), p-value< 0.0001; $\hat{\tau}^2_{DL} = 0.017$ and Q-statistic=65.77 with p-value=0.09. This means that the studies are not very heterogeneous, and therefore it will be interesting to see the performance of the retrospective Gombay sequential bootstrap test for REM in comparison with the standard methods which are based on fixed effect model. The data and results of the analysis are presented in Tables 8.3-8.8 in the Appendix.

Let ϕ be the true value of the estimates of the log-relative risk, $\hat{\phi}_i$. For a new intervention the objective is to test the null hypothesis of no effect of chewing gum, $H_0 : \phi = 0$. The CMA based on random-effects model of the data indicates a significant result (p-value=0.031) at trial 3; SMA indicates significant result (z-value of 3.23 is greater than the upper bound of 2.81) at trial 5. The penalized Z-test based on the adjustment factor of $\lambda = 2$ indicates significant result (test value of 1.92 is greater than $Z_{1-0.05} = 1.64$) at trial 7, while the CUSUM indicates a significant result at trial 5. For the retrospective Gombay sequential bootstrap test for REM, GDL, GH, GPM and GREML indicate a significant result at trial 7 with test values of 0.5615, 0.5615, 0.5668 and 0.5468, respectively. Thus we conclude that there is a significant effect of NRT.

To monitor for any possible trend in the effect over time, observe that in Figure 5.10 the CMA begins to show a gradual increase in effect from trial 5. So a new target value of 0.41 that corresponds to the combined log-relative risk at trial 5 is set. Observe that in Figures 5.10 and 5.12 only the CUSUM indicates a significant result at trial 38. However it is worth to note that the CUSUM test by its definition does not take into account the heterogeneity variance, τ^2 in random-effects model of meta-analysis, and thus the result is likely to be spurious.

5.4 Discussion of the use of Gombay method for sequential random-effects meta-analysis

One of the objectives in this thesis has been to find a suitable statistical method for monitoring temporal trends in effect sizes in random-effects meta-analysis. In the process, the use of the Gombay method which has solid statistical foundations with the advantage of applicability to parametric models in the presence of nuisance parameter was introduced in Chapter 4. However, using the Gombay method with standard critical values derived from asymptotic theory leads to a test with incorrect probability of Type I error rate. As commented by Horowitz (1997), '*Asymptotic theory often provides inaccurate approximation of the limiting distributions of test statistic, which can result in a test with different true and nominal levels*'. Therefore, this Chapter considered the use of bootstrap-based critical values with Gombay method for the sequential random-effects meta-analysis (retrospective Gombay sequential bootstrap test for REM). Simulations were conducted for the test based on four different estimators of τ^2; DerSimonian and Laird (1986), Higgins et al. (2011), Paule and Mandel (1982) and the REML. As have been shown, the bootstrap critical values provide a remarkable reduction in the difference between the true and the nominal level.

The Type I error rates achieved by the retrospective Gombay sequential bootstrap test for REM based on all the four estimators of τ^2 considered are close to the nominal level. The test based on DerSimonian and Laird (1986) and Higgins et al. (2011) estimators of τ^2 have more statistical power compared to the test based on Paule and Mandel (1982) and the REML when the number of studies is small and the reverse is the case when the number of studies is large. The retrospective Gombay sequential bootstrap test controls the Type I error better than the penalized Z-test and SMA. Unlike the penalized Z-test and SMA where the Type I errors vary for different values of K, n and τ^2, the Type I errors in Gombay test for REM with bootstrap critical values based on all the four estimators of τ^2 considered are relatively stable.

This Chapter also demonstrated the application of the Gombay method to odds ratio and relative risk effect sizes using two meta-analytic examples from medicine. As have been shown, the method allows sequential evaluation of treatment effect and monitoring of temporal trends in magnitude of effect sizes. Statistical significance of the treatment effect is established in both examples while temporal trends are detected in the first example, see Figures 5.9 and 5.12. The retrospective Gombay sequential bootstrap test is comparable to the other methods in terms of early detection of shifts in treatment effect. As have been seen, CMA, SMA and the CUSUM all detect change in magnitude of effect at stage 7. For the retrospective Gombay sequential bootstrap test, GDL and GH indicate significant shift in treatment effect at trial 6 while GPM and GREML at trial 5 in the first example. In the second example, the CUSUM and SMA detect change in magnitude of effect at stage 5, while the penalized Z-test and the retrospective Gombay sequential bootstrap tests based on all the four estimators of τ^2 at stage 7.

In many conventional sequential methods for meta-analysis, the between-study variation, τ^2 is not included in the determination of the sequential boundaries, and this often leads to the inflation of Type I error when the treatment effects are substantially heterogeneous. But in the retrospective Gombay sequential bootstrap test, the problem is taken care of as the τ^2 is included in the calculation of the bootstrap critical values. As can be seen, the retrospective Gombay sequential bootstrap test for REM controls the Type I error better compared to the penalized Z-test and SMA.

Calculation of the bootstrap critical values requires that the entire data be available at the start of the analysis, and therefore the application of the present method may be limited to retrospective meta-analysis. However meta-analysis is a quantitative approach for systematic assessment of the results of previous research in order to arrive at conclusions about the body of research (Petitti, 1999). Therefore, sequential methods in retrospective meta-analysis can be used to decide whether enough evidence has been gathered so that further trials are unnecessary. They can also be used for deciding whether an existing meta-analysis should be updated or not. The Gombay method with bootstrap critical values can be used prospectively when the maximum number of studies K required for the analysis is known in advance. Roloff et al. (2013), Kulinskaya and Wood (2014) discussed how the maximum number of studies required in a sequential meta-analysis can be determined in FEM and REM based on power analysis. Their methods can be utilised for this purpose.

In general, the Gombay method with bootstrap critical values controls the Type I error well irrespective of the number of studies, average sample size and the amount of heterogeneity in treatment effects. The method is comparable with standard sequential methods in meta-analysis in terms of allowing sequential evaluation of accumulating evidence and early detection of shifts in treatment effect. On the basis of the simulations, the DerSimonian and Laird (1986) and Higgins et al. (2011) estimators of τ^2 work well in the Gombay test for REM work well for small and medium number of studies, while Paule and Mandel (1982) and REML estimators are slightly better for large studies.

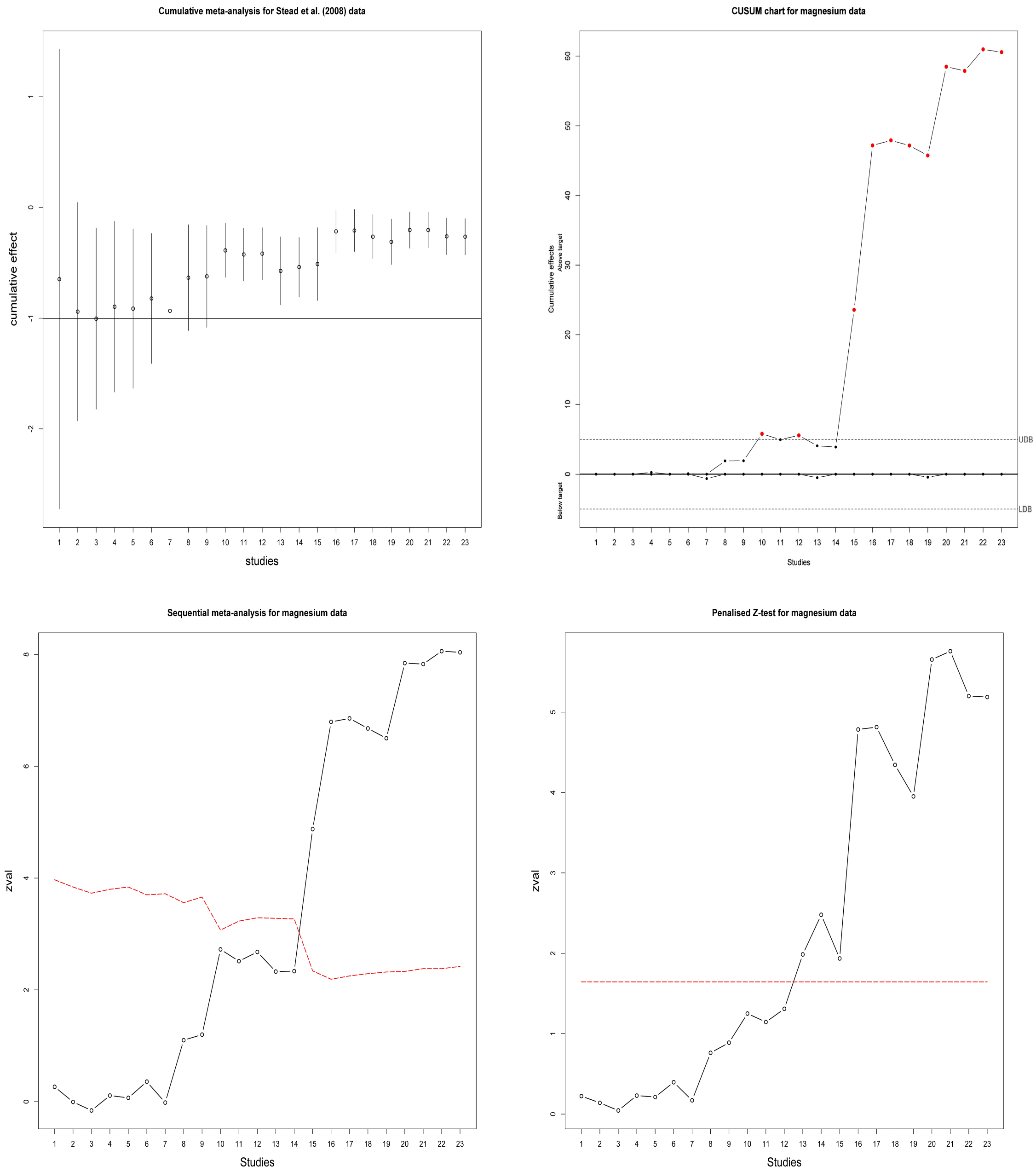


Figure 5.7: Analysis of magnesium for myocardial infarction (Li et al., 2007) data using Cumulative, CUSUM, Sequential meta-analysis and penalised Z-test for magnesium data. CMA and SMA are based on $\hat{\tau}^2_{DL}$. The horizontal line is the combined log odds ratio -0.934 (OR=0.393) at trial 7. The same value is the target value for SMA. The red dotted line is the upper-boundary value for the one-sided test which is first crossed at trial 13. The control limits for CUSUM chart (dashed lines) are defined at ±5σ. The red dashed line on the penalised Z-test plot is the one-sided upper boundary value.

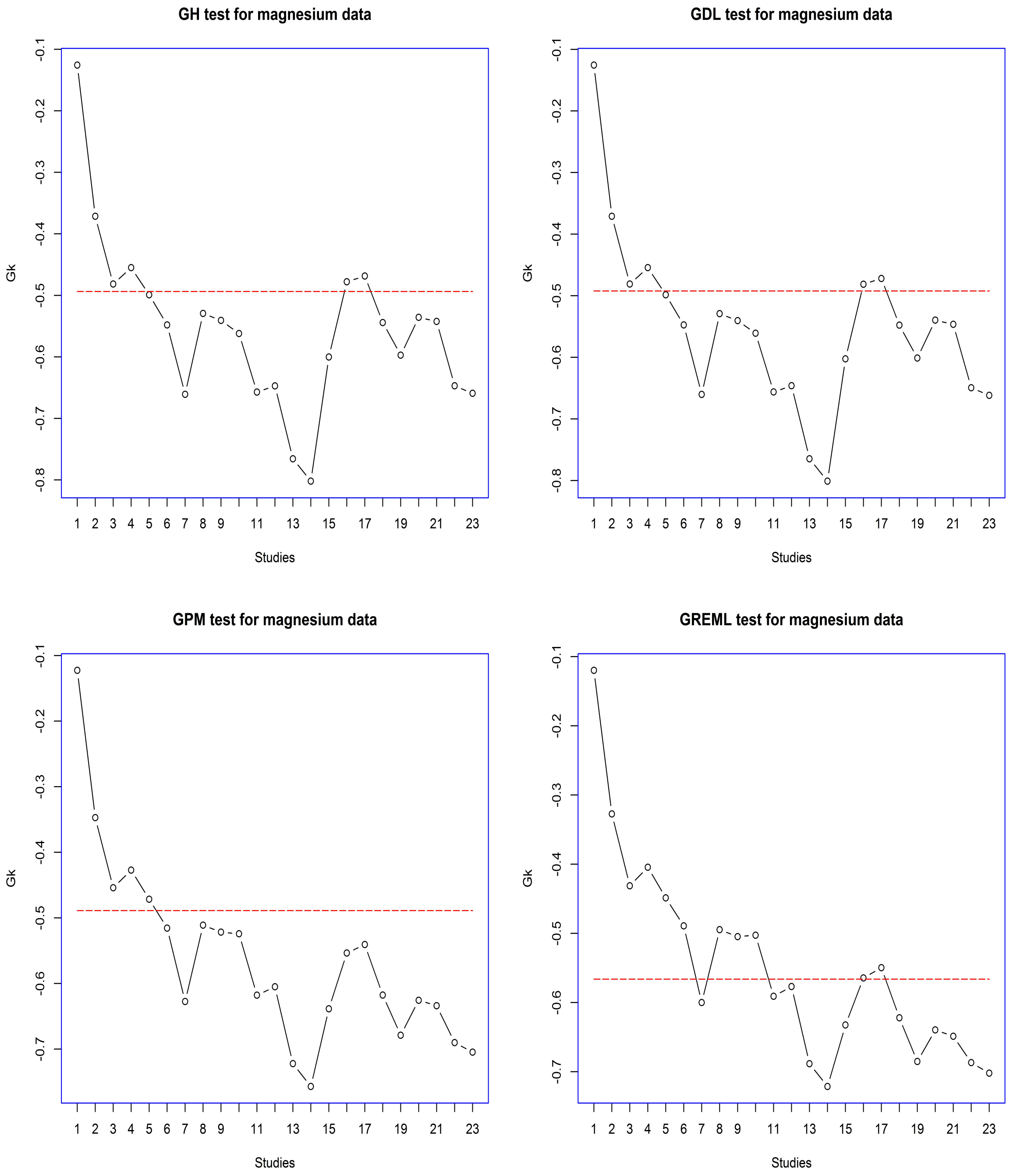


Figure 5.8: Analysis of magnesium for myocardial infarction (Li et al., 2007) data using the retrospective Gombay sequential bootstrap test for REM based on Higgins et al. (2011), DerSimonian and Laird (1986), Paule and Mandel (1982) and REML estimators of τ^2 (GDL, GH, GPM and GREML). The target value is set at 0, and the red dashed lines in GDL, GH, GPM and GREML plots are the one-sided lower boundary values.

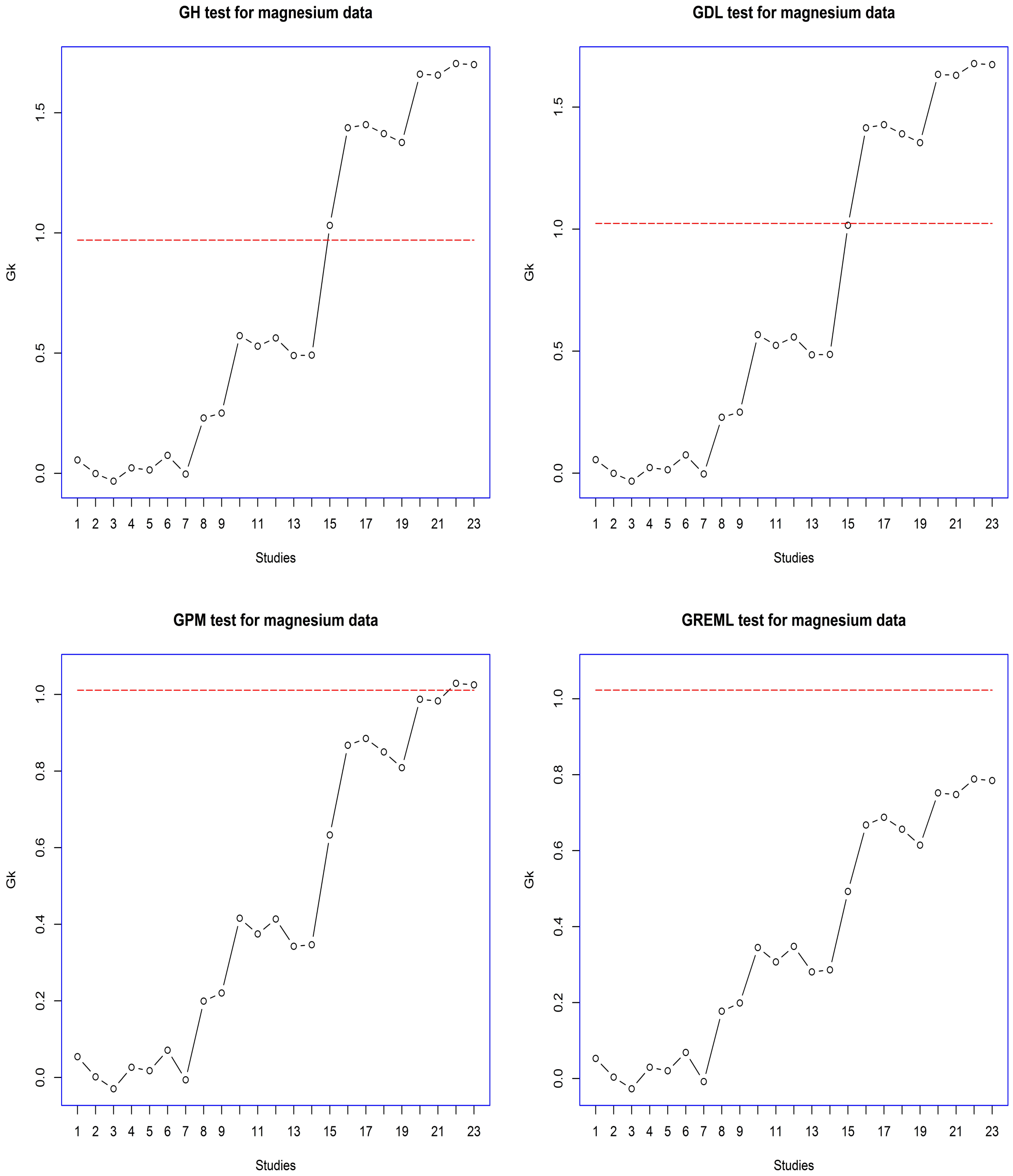


Figure 5.9: Analysis of magnesium for myocardial infarction (Li et al., 2007) data using the retrospective Gombay sequential bootstrap test for REM based on Higgins et al. (2011), DerSimonian and Laird (1986), Paule and Mandel (1982) and REML estimators of τ^2 (GDL, GH, GPM and GREML). The target value is set at -0.934, and the red dashed lines in GDL, GH, GPM and GREML plots are the one-sided upper boundary values.

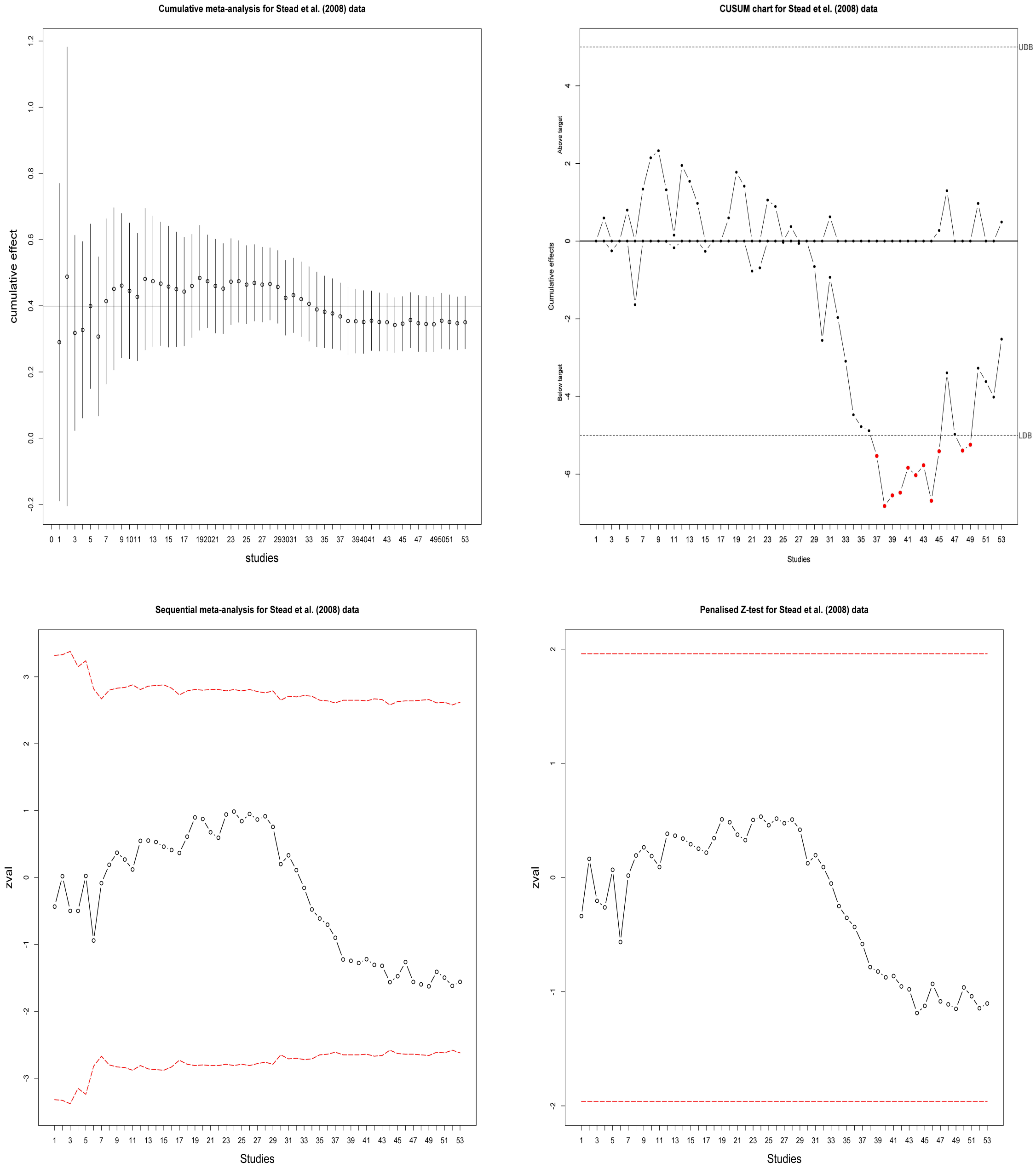
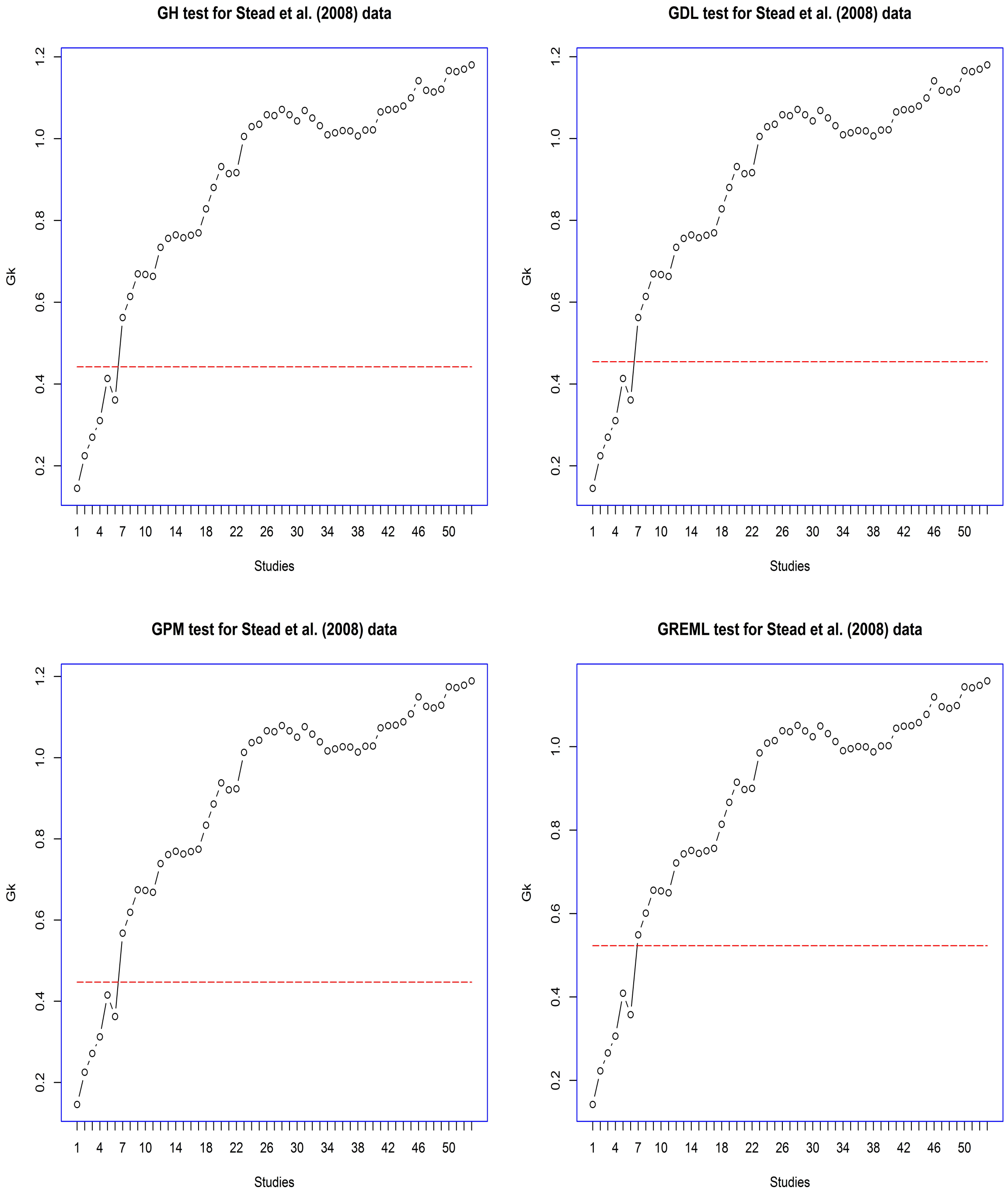


Figure 5.10: Analysis of Stead et al. (2008) data using Cumulative, CUSUM, Sequential meta-analysis and penalised Z-test. CMA and SMA are based on REM and $\hat{\tau}^2_{DL}$. The horizontal line is the combined log relative risk 0.41 (RR=1.51) at trial 5. The same value of 0.41 is used as the target value for SMA. The control limits for CUSUM charts are defined at $\pm 5\sigma$. The red dotted lines on the SMA plots and penalised Z-test are the upper and lower boundary values for two-sided tests.


GH test for Stead et al. (2008) data

GDL test for Stead et al. (2008) data

GPM test for Stead et al. (2008) data

GREML test for Stead et al. (2008) data

Figure 5.11: Analysis of Stead et al. (2008) data using the retrospective Gombay sequential bootstrap test for REM based on Higgins et al. (2011), DerSimonian and Laird (1986), Paule and Mandel (1982) and REML estimators of τ^2 (GDL, GH, GPM and GREML). The target value is set at 0 and the red dashed lines in GDL, GH, GPM and GREML tests plots are the upper boundary values for one-sided tests.

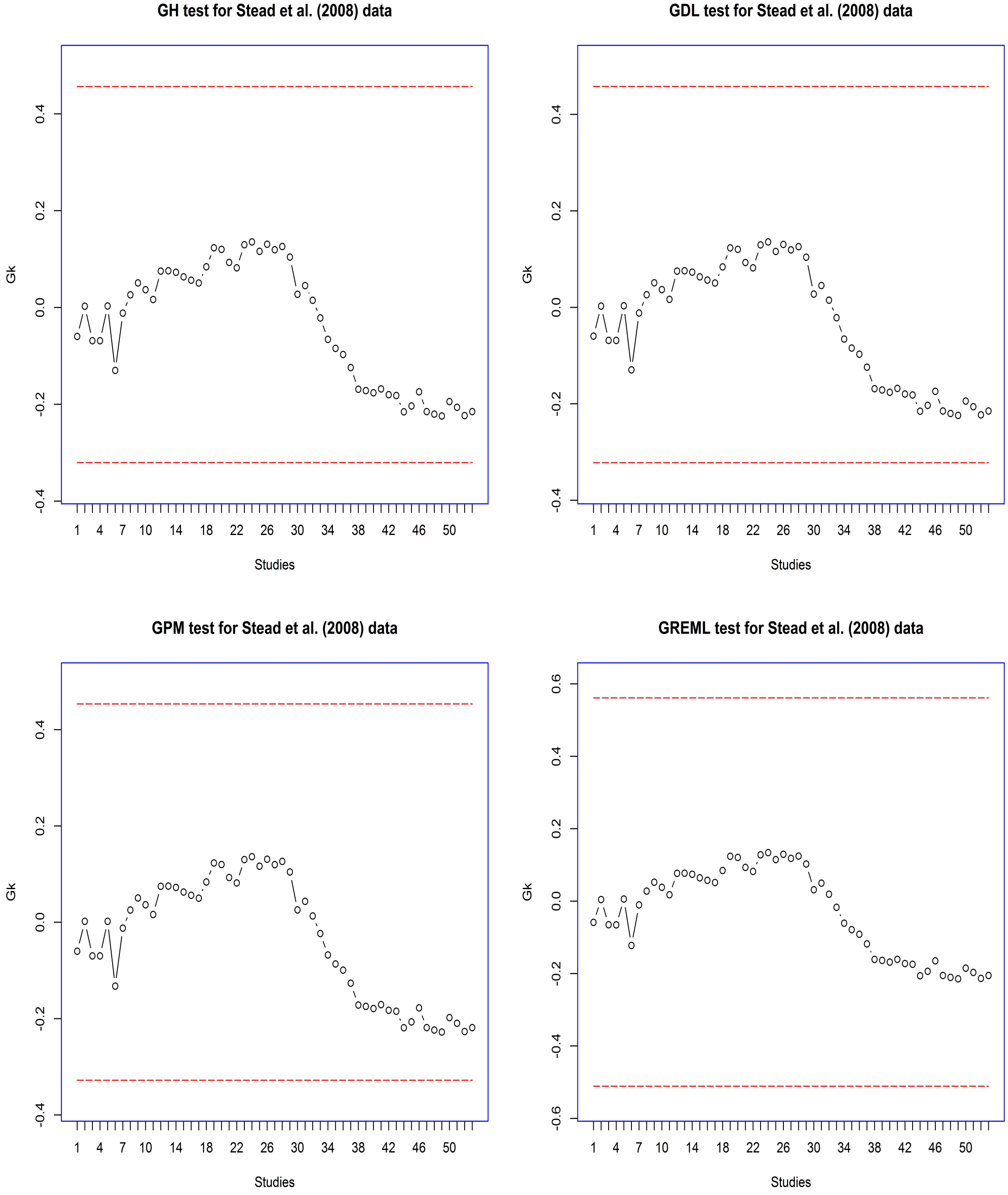


Figure 5.12: Analysis of Stead et al. (2008) data using the retrospective Gombay sequential bootstrap test for REM based on Higgins et al. (2011), DerSimonian and Laird (1986), Paule and Mandel (1982) and REML estimators of τ^2 (GDL, GH, GPM and GREML). The target value is set at 0.41 and the double red dashed lines in GDL, GH, GPM and GREML tests plots are the lower and upper boundary values for two-sided tests.

Chapter 6

Sequential bias in accumulating evidence.

The effect of existing evidence in meta-analysis was introduced in Section 1.3 of Chapter 1. Two ways of using existing evidence to inform further research, sequential decision and sequential design, were identified. This Chapter discusses the bias arising in these methods. Specifically, Section 6.1 discusses sequential decision bias, models for probability of running the next trial and simulations on sequential decision bias. Section 6.2 discusses sequential design bias and simulations on the sequential design bias. In Section 6.3, the application of sequential decision and sequential design bias is provided using an example of a meta-analysis by Johnson (1993). Section 6.4 is discussion of sequential bias in accumulating evidence.

6.1 Sequential decision bias

As defined in Section 1.3, sequential decision is a method of using existing information in making decision to conduct a new trial. The next sections discuss the bias associated with this method. It is assumed throughout this Section that the estimates of an effect $\hat{\theta}_i \sim N(\theta, \sigma_i^2)$.

6.1.1 Bias derivation

To illustrate sequential decision bias, consider the following simple situation. Suppose there is a study which had estimated the effect of interest, θ, by $\hat{\theta}_1$ and its variance by s_1^2. A researcher is considering the usefulness of running another study. Suppose that the probability of running this new study $p_1 = p(\hat{\theta}_1, S_1^2, \theta)$ is a function of the estimated effect and the effect of clinical interest θ_0, that is the same in both studies. If the second study is conducted, denote by ω_i the normalized inverse variance weights for $\hat{\theta}_i$, that is, $\omega_1 + \omega_2 = 1$. Then the combined effect is

$$\hat{\theta}_{(2)} = \begin{cases} \omega_1\hat{\theta}_1 + \omega_2\hat{\theta}_2, & \text{with probability } p(\hat{\theta}_1, S_1^2, \theta_0) \\ \\ \hat{\theta}_1, & \text{with probability } 1 - p(\hat{\theta}_1, S_1^2, \theta_0) \end{cases} \tag{6.1}$$

Assuming that $\hat{\theta}_1$ and $\hat{\theta}_2$ are independent, and that the weights are either non-random, as is common to assume in meta-analysis, or at least independent of the estimated effects, which is true for the means of the continuous outcomes and the usual weights based on inverse sample variances, the expected value of the combined effect is

$$\mathrm{E}[\hat{\theta}_{(2)}] = \mathrm{E}\left\{p(\hat{\theta}_1, S_1^2, \theta_0)(\omega_1\hat{\theta}_1 + \omega_2\hat{\theta}_2) + (1 - p(\hat{\theta}_1, S_1^2, \theta_0))\hat{\theta}_1\right\}. \tag{6.2}$$

Substituting $1 - \omega_1$ for ω_2 equation (6.2) results to

$$\begin{aligned}
\mathrm{E}[\hat{\theta}_{(2)}] =& \mathrm{E}\left\{p(\hat{\theta}_1, S_1^2, \theta_0)\omega_1\hat{\theta}_1 + p(\hat{\theta}_1, S_1^2, \theta_0)(1-\omega_1)\hat{\theta}_2 + (1 - p(\hat{\theta}_1, S_1^2, \theta_0))\hat{\theta}_1\right\} \\
=& \mathrm{E}\left\{\mathrm{E}\left\{p(\hat{\theta}_1, S_1^2, \theta_0)\omega_1\hat{\theta}_1 + p(\hat{\theta}_1, S_1^2, \theta_0)(1-\omega_1)\hat{\theta}_2 + (1 - p(\hat{\theta}_1, S_1^2, \theta_0))\hat{\theta}_1|\hat{\theta}_1\right\}\right\} \\
=& \mathrm{E}\left\{p(\hat{\theta}_1, S_1^2, \theta_0)(1-\omega_1) + \hat{\theta}_1 - p(\hat{\theta}_1, S_1^2, \theta_0)(1-\omega_1)\hat{\theta}_1\right\} \\
=& \theta + (\omega_1 - 1)\mathrm{cov}(p(\hat{\theta}_1, S_1^2, \theta_0), \hat{\theta}_1).
\end{aligned} \tag{6.3}$$

Assume that the expected value of the probability, $\mathrm{E}(p(\hat{\theta}_1, \theta_0)) \neq 0$ and let $Y = 1$ if the second trial is conducted and zero otherwise. Suppose Y and $\hat{\theta}_2$ are conditionally independent given $\hat{\theta}_1$. Then the conditional expectation given the second trial is conducted is given by

$$\begin{aligned}
\mathrm{E}\left\{\hat{\theta}_{(2)}\middle|Y=1\right\} =& \frac{\mathrm{E}\left\{(\omega_1\hat{\theta}_1 + \omega_2\hat{\theta}_2)Y\right\}}{p(Y=1)} \\
=& \frac{\mathrm{E}\left\{\mathrm{E}\left\{(\omega_1\hat{\theta}_1 + \omega_2\hat{\theta}_2)Y|\hat{\theta}_1\right\}\right\}}{p(Y=1)} \\
=& \frac{\mathrm{E}\left\{\omega_1\hat{\theta}_1 p(\hat{\theta}_1, S_1^2, \theta_0) + (1+\omega_1)\theta p(\hat{\theta}_1, S_1^2, \theta_0)\right\}}{p(Y=1)} \\
=& \frac{\theta\mathrm{E}(p(\hat{\theta}_1, S_1^2, \theta_0)) + \omega_1\mathrm{E}(\hat{\theta}_1 p(\hat{\theta}_1, S_1^2, \theta_0) - \theta p(\hat{\theta}_1, S_1^2, \theta_0))}{\mathrm{E}(p(\hat{\theta}_1, \theta_0))} \\
=& \theta + \frac{\omega_1\mathrm{cov}(p(\hat{\theta}_1, S_1^2, \theta_0), \hat{\theta}_1)}{\mathrm{E}(p(\hat{\theta}_1, S_1^2, \theta_0))} + \omega_1\left\{\mathrm{E}(\hat{\theta}_1 - \theta)\right\}.
\end{aligned} \tag{6.4}$$

Thus, unless $\mathrm{cov}(p(\hat{\theta}_1, S_1^2, \theta_0), \hat{\theta}_1) = 0$, both the unconditional and conditional expectations are biased. The last term in the above equation, though zero for an unbiased estimator $\hat{\theta}_1$, is retained intentionally, so that the equation can be generalized to the case of K sequential decisions and trials. Similarly, the conditional expectation given that the second trial is not conducted is given by

$$\begin{aligned}\mathrm{E}\left\{\hat{\theta}_{(2)}\middle|Y=0\right\} =&\mathrm{E}(\hat{\theta}_1\middle|Y=0)\\ =&\frac{\mathrm{E}\left\{\hat{\theta}_1(1-Y)\right\}}{p(Y=0)}\\ =&\frac{\mathrm{E}\left\{\hat{\theta}_1(1-\mathrm{E}(Y|\hat{\theta}_1))\right\}}{p(Y=0)}\\ =&\frac{\mathrm{E}\left\{\hat{\theta}_1(1-p(\hat{\theta}_1,S_1^2,\theta_0)))\right\}}{p(Y=0)}\\ =&\theta-\frac{\mathrm{cov}(p(\hat{\theta}_1,S_1^2,\theta_0),\hat{\theta}_1)}{1-\mathrm{E}(p(\hat{\theta}_1,S_1^2,\theta_0))}.\end{aligned} \tag{6.5}$$

Remark 6.1.1. *Suppose K trials were run sequentially, and the decision to run trial $i+1$ was dependent on the cumulative results from the first i trials, $\hat{\theta}_{(i)}=\sum_{j=1}^{j}\omega_j\hat{\theta}_j$ for $i=1,\ 2,\ ...,\ K-1$. Equation 6.3 can be applied directly to cumulative effect $\hat{\theta}_{(K-1)}$ and the effect in the $K-th$ trial $\hat{\theta}_K$, to obtain a recurrent equation for sequential decision bias*

$$E_K(\hat{\theta}_K-\theta)=\omega_{(K-1)}E_{K-1}(\hat{\theta}_{(K-1)}-\theta)+\omega_{K+1}cov(p_{(K-1)},S^2_{(K-1)},\hat{\theta}_{(K-1)})\left[E(p_{(K-1)})\right]^{-1}. \tag{6.6}$$

In equation 6.6, $E_i(.)$ is the conditional expectation given i trials, and $\omega_{(i)}=\sum_1^i\omega_j/\sum_1^{i+1}\omega_j$ is the normalized weight for $\hat{\theta}_{(i)}$. Similarly, $p_{i-1}=p(\hat{\theta}_{(i-1)},S^2_{(i-1)},\theta_0)$ is the probability of running the $i-th$ trial.

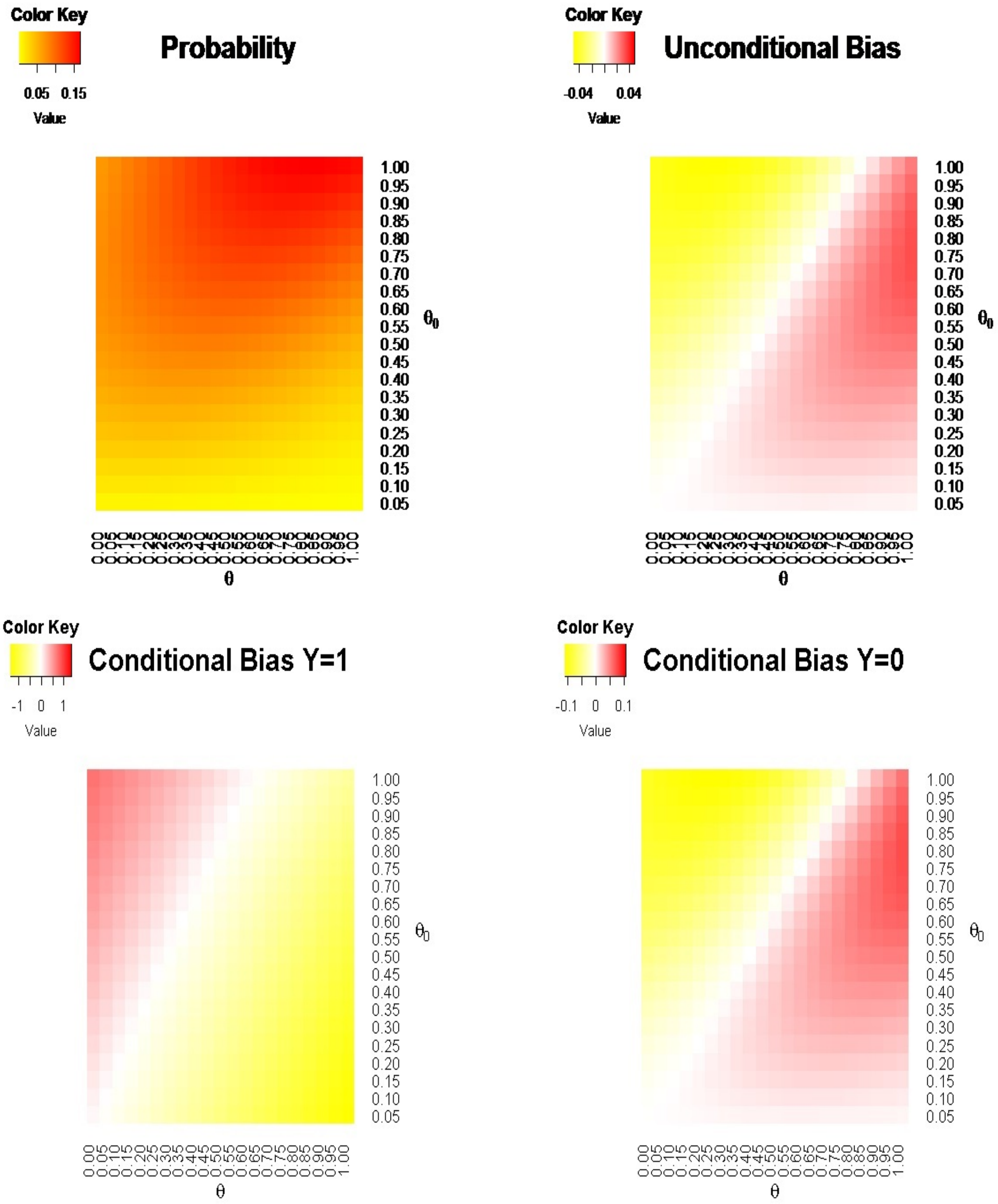
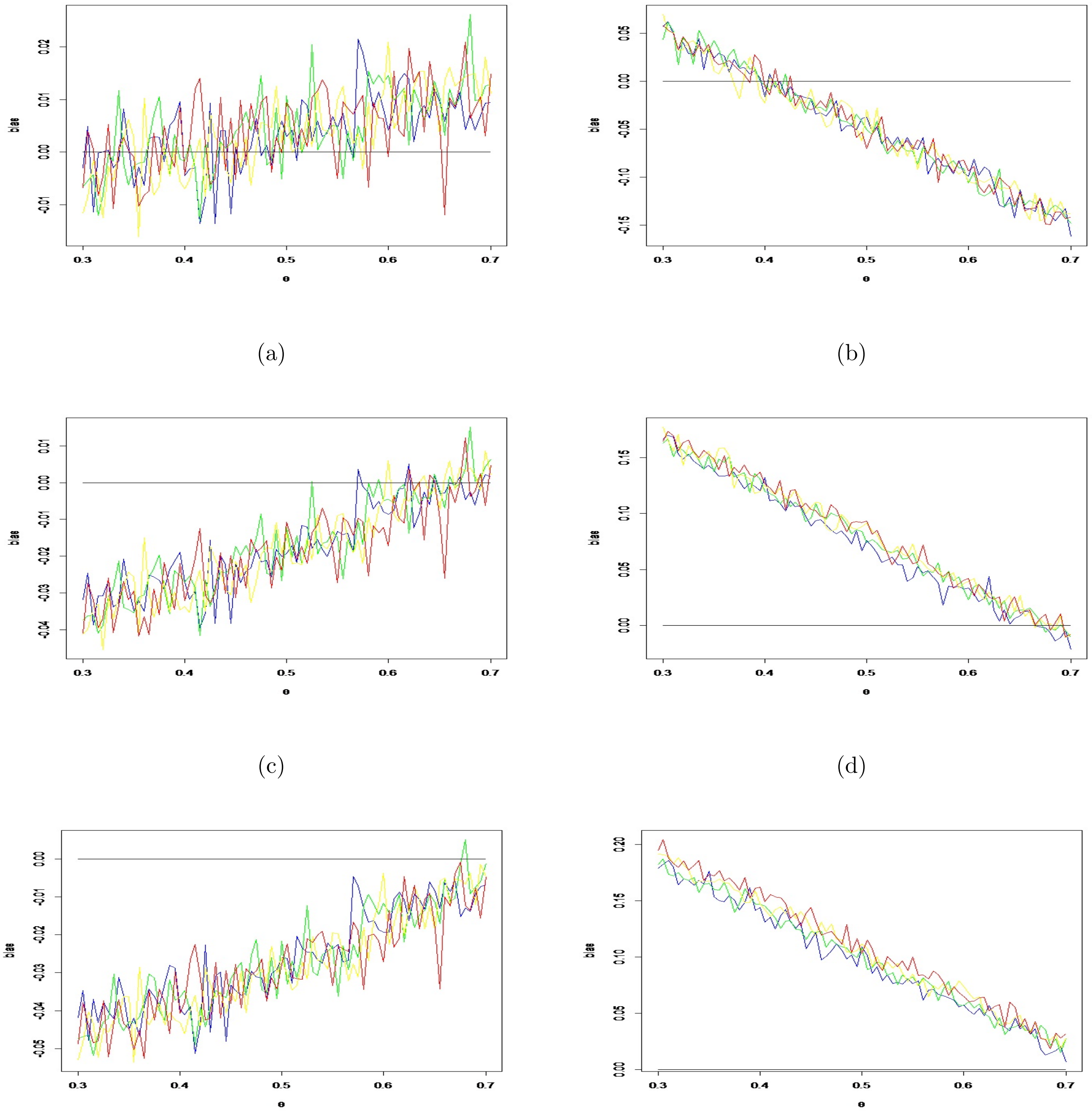


Figure 6.1: Probability of conducting a second trial and the bias in the unconditional and conditional means when the second trial is conducted (Y=1) and not conducted (Y=0), given by equations (1), (2) and (3), respectively. The X-axis is the true value of θ (effect parameter) while the Y-axis is the target value θ_0. The parameter values are $\sigma = 0.2$, $t = 3$, and $\omega_1 = \omega_2 = 1/2$.

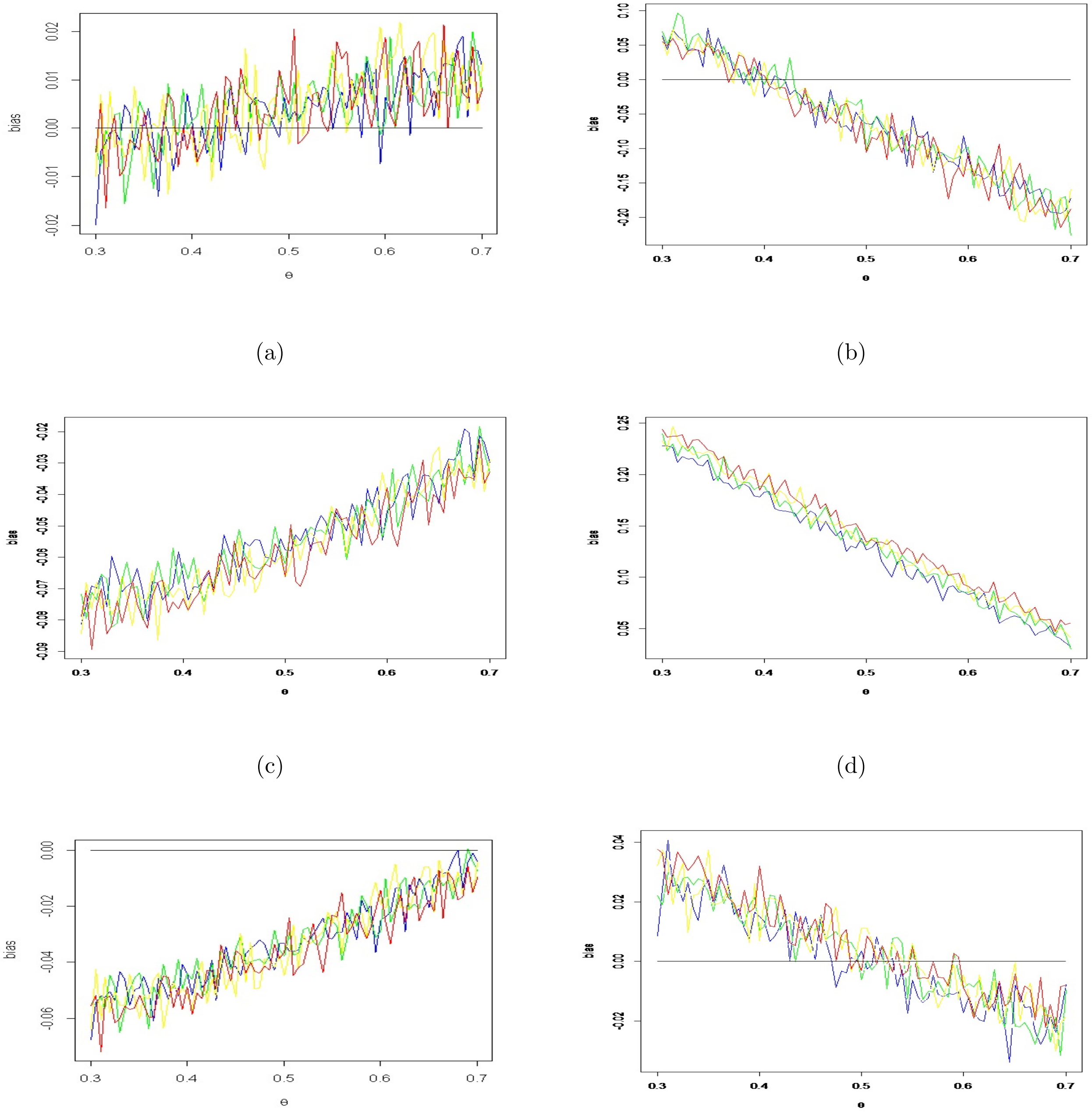


(a) (b)

(c) (d)

(e) (f)

Figure 6.2: Biases of unconditional (left) and conditional $Y = 1$ (right) expected values of the cumulative effects $\hat{\theta}_{(2)}$ at the second study for τ^2 values of 0 (blue), 0.02 (green), 0.04 (yellow) and 0.06 (red). The rows 1 to 3 correspond to the biases in power-law with $t = 3$, extreme value ($r = 0.8$) and probit ($r = 0.8$)models with $\alpha = 0$ and $\beta = 1$, respectively. Results from 10000 simulations at each value of $\theta = 03(0.05)0.7$ for the target value of $\theta_0 = 0.5$, equal weights $\omega_1 = \omega_2 = \omega_3$ and the variance $\sigma^2 = 1/3$ (corresponding to the within-study variance $s_1^2 = 19.94$ and sample size of $n_1 = 61$ in Example of Section 6.3).



(a)

(b)

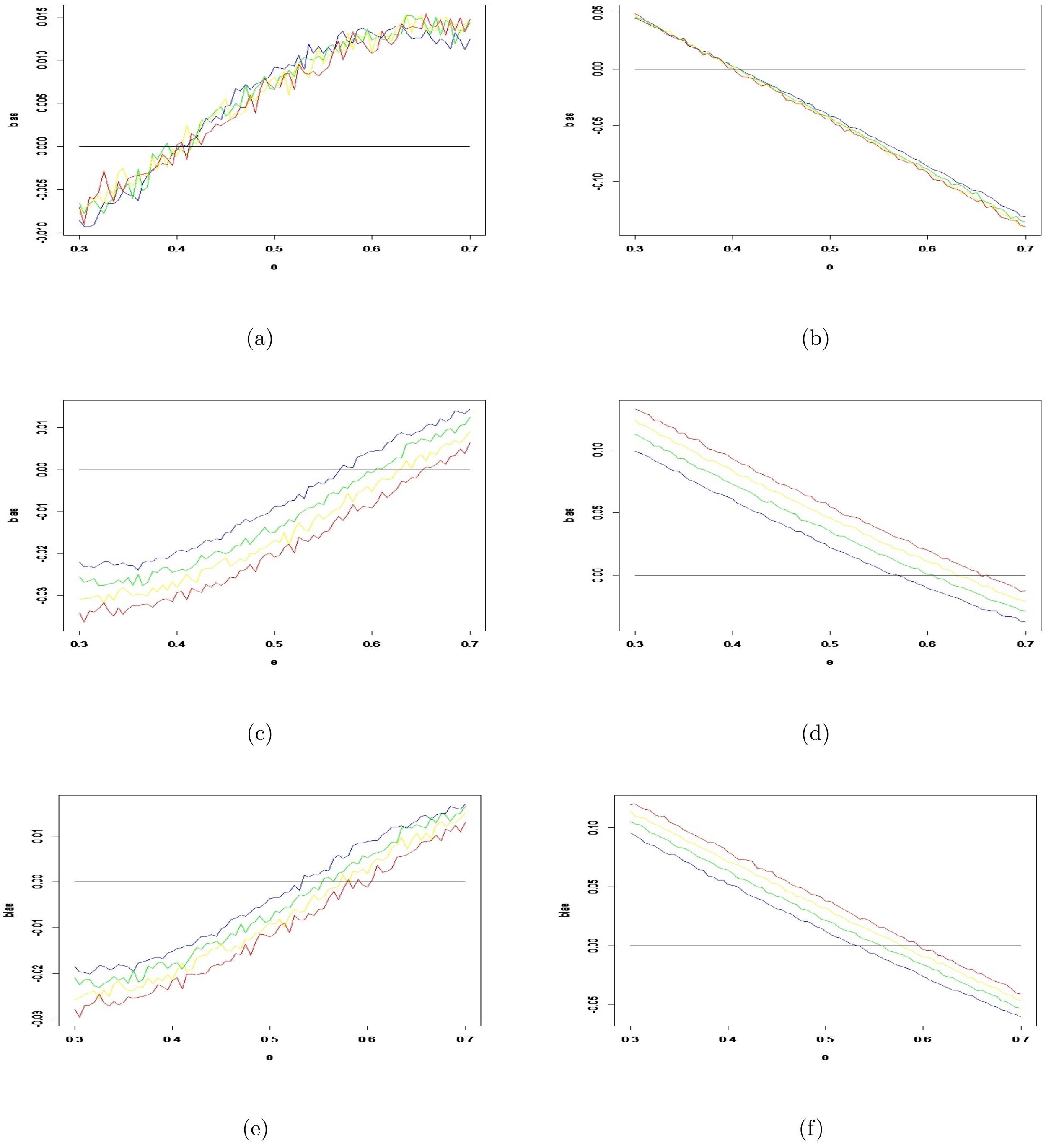
(c)

(d)

(e)

(f)

Figure 6.3: Biases of unconditional (left) and conditional $Y = 1$ (right) expected values of the cumulative effects $\hat{\theta}_{(3)}$ at the second study for τ^2 values of 0 (blue), 0.02 (green), 0.04 (yellow) and 0.06 (red). The rows 1 to 3 correspond to the biases in power-law with $t = 3$, extreme value ($r = 0.8$) and probit ($r = 0.8$) models with $\alpha = 0$ and $\beta = 1$, respectively. Results from 10000 simulations at each value of $\theta = 03(0.05)0.7$ for the target value of $\theta_0 = 0.5$, equal weights $\omega_1 = \omega_2 = \omega_3$ and the variance $\sigma^2 = 1/3$ (corresponding to the within-study variance $s_1^2 = 19.94$ and sample size of $n_1 = 61$ in Example of Section 6.3).



(a)

(b)

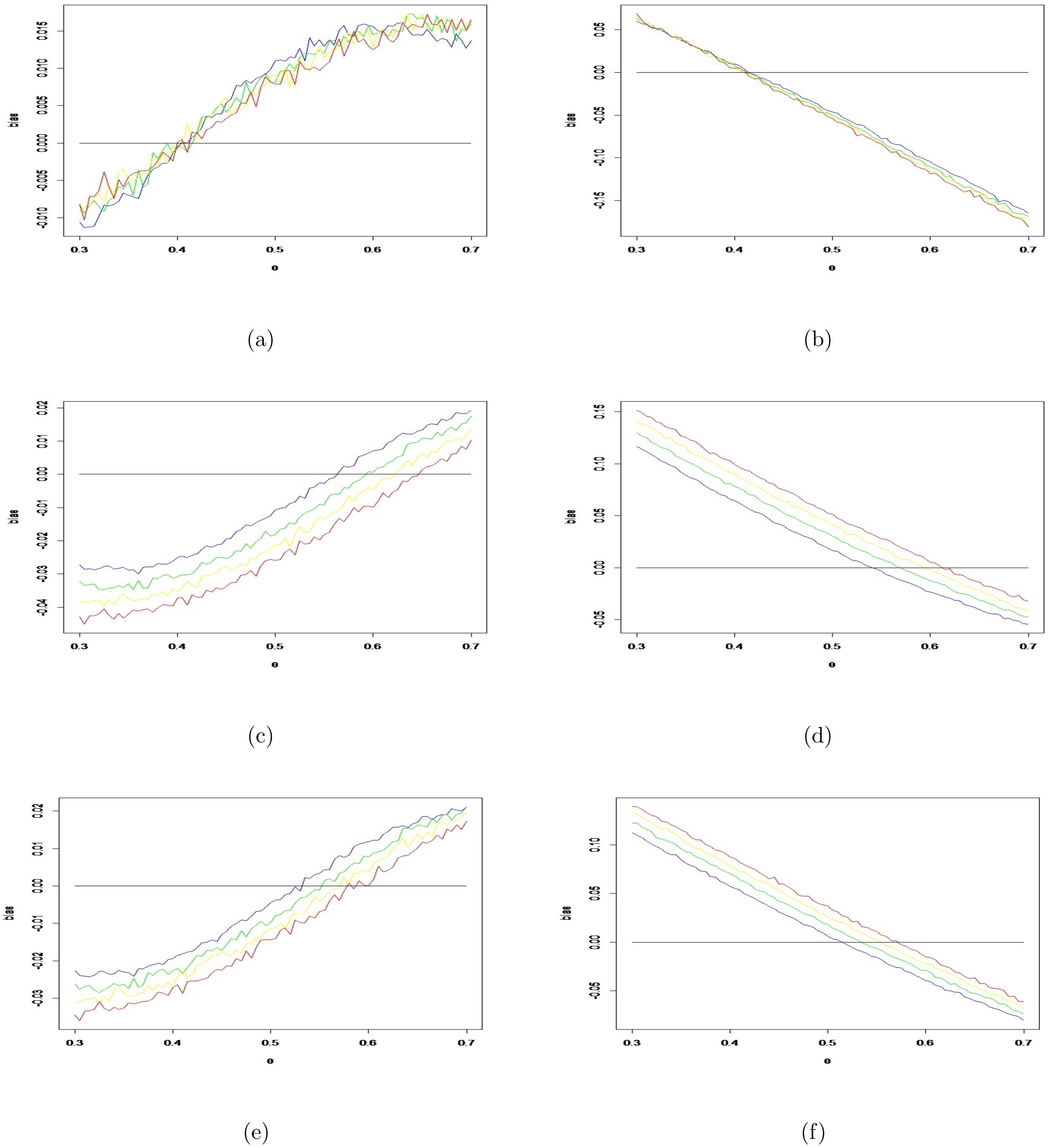
(c)

(d)

(e)

(f)

Figure 6.4: Biases of unconditional (left) and conditional $Y = 1$ (right) expected values of the cumulative effects $\hat{\theta}_{(2)}$ at the second study for τ^2 values of 0 (blue), 0.02 (green), 0.04 (yellow) and 0.06 (red). The rows 1 to 3 correspond to the biases in power-law with $t = 3$, extreme value ($r = 0.8$) and probit ($r = 0.8$) models with $\alpha = 0$ and $\beta = 1$, respectively. Results from 100000 simulations at each value of $\theta = 03(0.05)0.7$ for the target value of $\theta_0 = 0.5$, equal weights $\omega_1 = \omega_2 = \omega_3$ and the variance $\sigma^2 = 1/3$ (corresponding to the within-study variance $s_1^2 = 19.94$ in example of Section 6.3) and sample size of $n_1 = 500$.



(a) (b)

(c) (d)

(e) (f)

Figure 6.5: Biases of unconditional (left) and conditional $Y = 1$ (right) expected values of the cumulative effects $\hat{\theta}_{(3)}$ at the second study for τ^2 values of 0 (blue), 0.02 (green), 0.04 (yellow) and 0.06 (red). The rows 1 to 3 correspond to the biases in power-law with $t = 3$, extreme value ($r = 0.8$) and probit ($r = 0.8$) models with $\alpha = 0$ and $\beta = 1$, respectively. Results from 100000 simulations at each value of $\theta = 03(0.05)0.7$ for the target value of $\theta_0 = 0.5$, equal weights $\omega_1 = \omega_2 = \omega_3$ and the variance $\sigma^2 = 1/3$ (corresponding to the within-study variance $s_1^2 = 19.94$ in example of Section 6.3) and sample size of $n_1 = 500$.

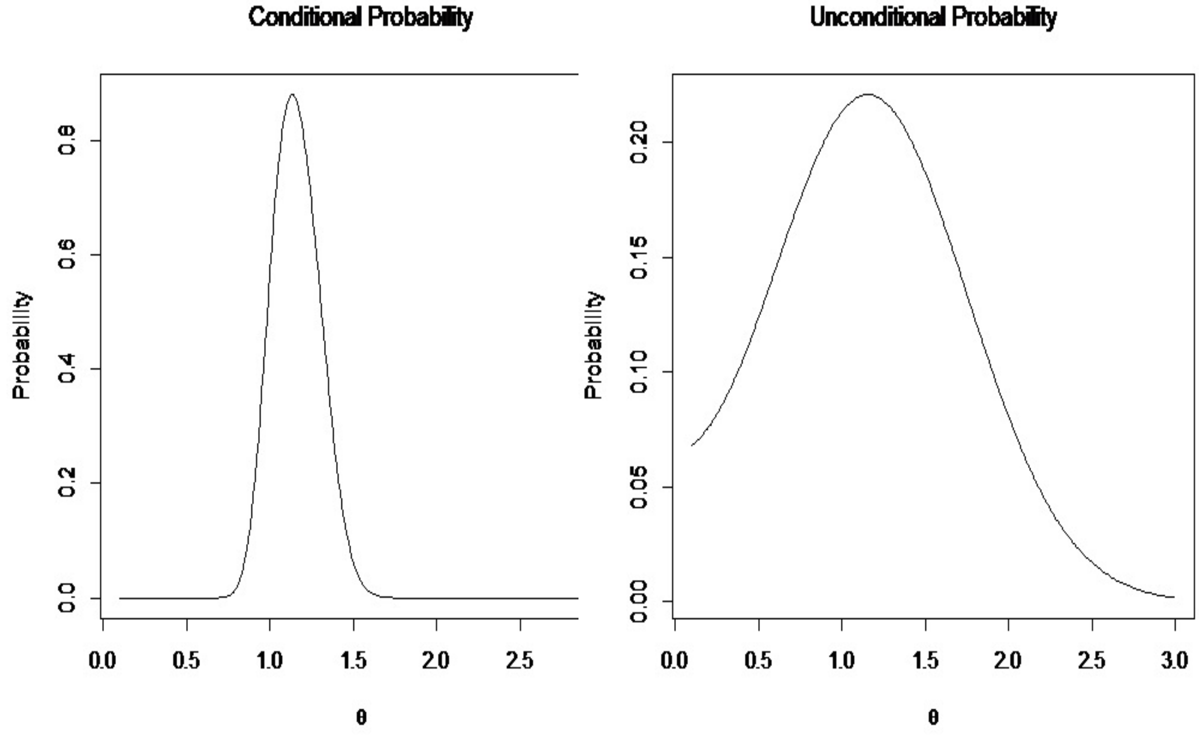


Figure 6.6: The conditional probability given that $\hat{\theta}_1 = \theta$, and the unconditional probability as a function of the true value θ of θ_0 of conducting the second trial for the power calculation rule of equations (6.17) and (6.18) with $n_1 = 60$, $\sigma_1^2 = 19.9$, $a = 30$ and $b = 80$.

6.1.2 Models for probability of running the next trial $\Pr(\hat{\theta}, \theta_0)$

To estimate the resultant biases numerically, a model for the probability of running the next trial, $\Pr(\hat{\theta}, S_1^2, \theta_0)$ is required. This Section first introduces three simple models namely; power-law, extreme value and probit models, and then a more complex model based on power calculation in Section 6.1.2.4. Each of the models differs in terms of definition of the probability of conducting the next trial, and thus may result in a different

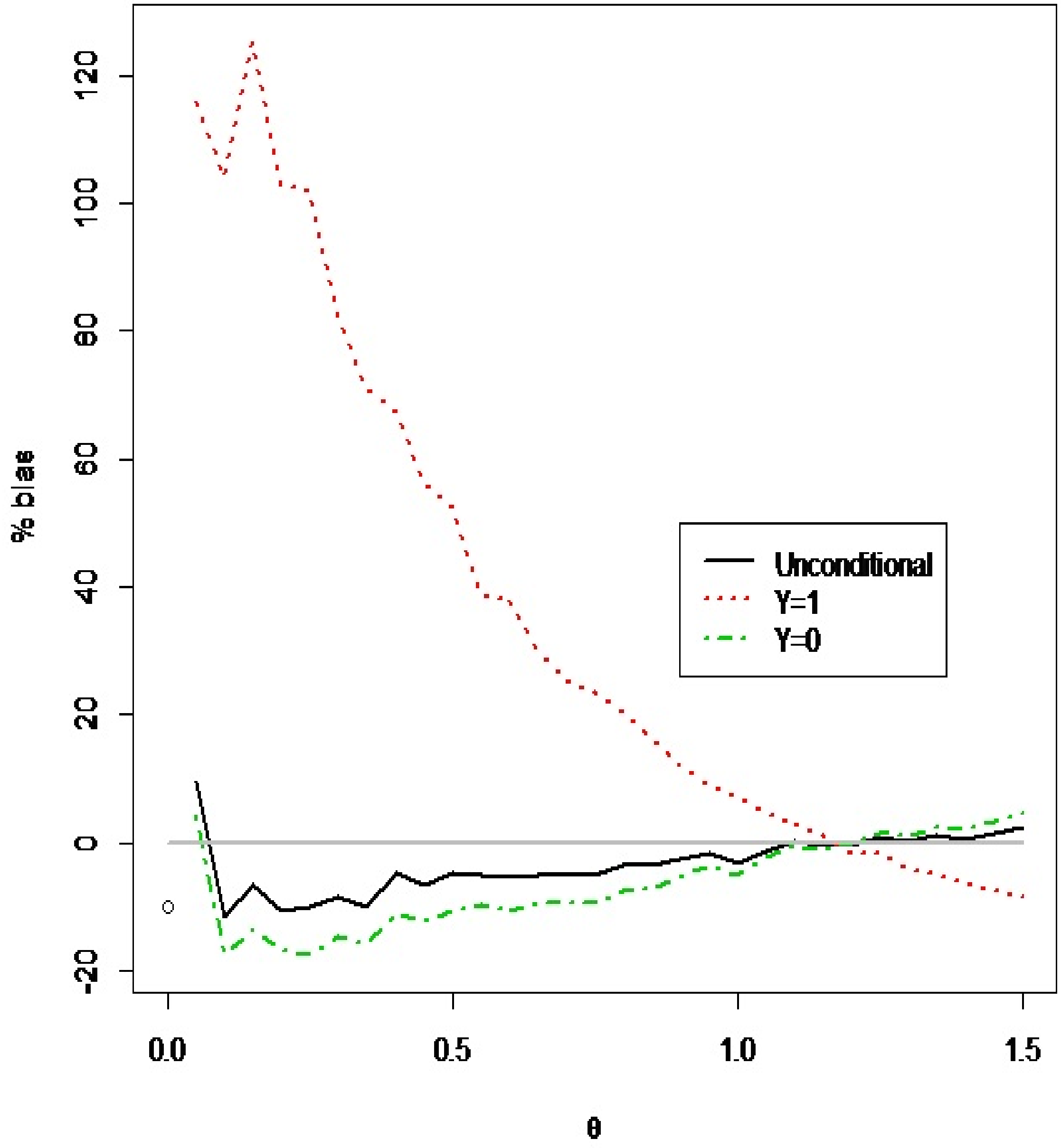


Figure 6.7: Estimated percent unconditional (solid line) and conditional (Y=1 dashed line, Y=0 dot-dashed line) bias from 10000 simulations as a function of θ for the power calculation rule of equations (6.17) and (6.18), with σ_1^2 estimated by the sample variance of the first trial of the example in the next Section 6.3, $S_1^2 = 19.99$, sample size $n_1 = 60$, $a = 30$, $b = 80$ and a second trial of size 80 is conducted if $a < n_2 < b$.

value of the bias. This is investigated empirically using simulations of unconditional and conditional means of the combined effect at stages 2 and 3 based on the power-law, extreme value and probit models in Section 6.1.2.3.

6.1.2.1 Power-law probability model

At the initial stages of clinical trials, if initial results are “promising” i.e. the combined estimate of results from studies is close to the effect of clinical interest θ_0, the trial is more likely to be continued. But if the initial results are not “promising” the combined estimate of results from studies will be significantly different from the target value and the trial may be stopped. Therefore in the power-law model, “promising” results increase the probability of running the next trial, and no further trials are required when the effect is at least θ_0. For $\theta_0 > 0$ and some $t > 0$, the power-law model for the probability of next trial is defined by

$$\Pr(\hat{\theta}, S^2, \theta_0) = \begin{cases} (\hat{\theta}/\theta_0)^t & \text{for } 0 < \hat{\theta} < \theta_0 \\ 0 & \text{elsewhere} \end{cases} \tag{6.7}$$

The variability of the estimator S^2 is not taken into account in this model, and thus it is a very simplistic model. The function $\mathrm{P}(\hat{\theta}, S^2, \theta_0) = \Pr(\hat{\theta}, \theta_0)$ is a distribution function from the power-law family of distributions on $[0,\, \theta_0]$ and $t = 1$ corresponds to uniform distribution. To evaluate the expected values, $\mathrm{E}[\Pr(\hat{\theta}, \theta_0)]$, $\mathrm{E}[\hat{\theta}, \Pr(\hat{\theta}, \theta_0)]$ and the covariance $\mathrm{cov}[\hat{\theta}, \Pr(\hat{\theta}, \theta_0)]$, the following integrals are required.

$$\begin{aligned}
\mathrm{E}\left\{\Pr(\hat{\theta}, \theta_0)\right\} &= \int\limits_0^{\theta_0} \left(\hat{\theta}/\theta_0\right)^t \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(\hat{\theta}-\theta)^2/(2\sigma^2)} d\hat{\theta} \\
\mathrm{E}\left\{\hat{\theta}\Pr(\hat{\theta}, \theta_0)\right\} &= \theta_0 \int\limits_0^{\theta_0} \left(\hat{\theta}/\theta_0\right)^{t+1} \frac{1}{2\pi\sigma^2} e^{-(\hat{\theta}-\theta)^2/(2\sigma^2)} d\hat{\theta} \qquad (6.8) \\
\mathrm{cov}\left(\hat{\theta}, \Pr(\hat{\theta}, \theta_0)\right) &= \int\limits_0^{\theta_0} \left(\hat{\theta}/\theta_0\right)^t (\hat{\theta}-\theta) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(\hat{\theta}-\theta)^2/(2\sigma^2)} d\hat{\theta}
\end{aligned}$$

From these equations, it can be seen that the covariance between $\hat{\theta}$ and $\Pr(\hat{\theta}, \theta_0)$ is negative if $\hat{\theta} > \theta_0$, so a negative bias for the conditional expectation of the probability of conducting the next trial is expected in this case. If $\theta_0 > \theta$ the bias is positive. The heat maps in Figure 6.1 show the bias for the unconditional and conditional mean as a function of $0 \leq \theta$, $\theta_0 \leq 1$ for $t = 3$, $\omega_1 = \omega_2 = 1/2$ and variance $\sigma^2 = 1/3$ (the value obtained as S_1^2/n_1 from the first trial in the example in Section 6.3). The unconditional and conditional biases given that the second trial is not conducted are positive when $\theta_0 > \theta$, while the reverse is the case for the conditional bias given that the second trial is conducted. These biases are investigated empirically using simulations in Section 6.1.2.3.

6.1.2.2 Extreme value and probit models

These models are derived based on a class of t-models for publication bias by Copas (2013). The models are of the form $a(\theta/\sigma)$ for an arbitrary function $a(.)$. The general form of these models is defined by

$$\Pr(\hat{\theta}, S^2, \theta_0, r) = \begin{cases} [1 - G((\theta - \theta_0)/\sigma)]/[1 - G((r-1)\theta_0/\sigma)] & \text{for } 0 < r(\theta) \\ \\ 0 & \text{elsewhere,} \end{cases} \tag{6.9}$$

for a distribution function $G(.)$. For the extreme value model, the distribution function $G(.)$ is given by

$$G(\theta, \sigma^2, \theta_0) = \exp(-\exp((\theta_0)/\sigma)), \tag{6.10}$$

and for probit model

$$G(\theta, \sigma^2, \theta_0) = \Phi(\alpha + \beta(\theta - \theta_0)/\sigma). \tag{6.11}$$

6.1.2.3 Empirical investigation of biases of the unconditional and conditional means based on power-law, extreme value and the probit models.

The simulations in this Section investigate the biases of unconditional and conditional means for the combined effects at stages 2 and 3, $\hat{\theta}_{(2)}$ and $\hat{\theta}_{(3)}$ based on power-law, extreme value and probit probability models using simulations. The effect of the variance component τ^2 and studies sizes on the bias were investigated. The simulations were designed as follows.

For $j = 1, ..., \text{K}$, where K is the number of simulations.

1. Simulate two effect size estimates $\hat{\theta}_{1j}$ and $\hat{\theta}_{2j}$ from normal distribution, $N(\theta, \tau^2 + \sigma^2/n)$ for studies 1 and 2, and let ω_1 and ω_2 be the normalised weights assigned to the studies 1 and 2, respectively.

2. Compute $\bar{\bar{p}}_{1j} = \Pr(\hat{\theta}_{1j}, S^2_{1j}, \theta_0)$ the probability of conducting the second trial, and use $\bar{\bar{p}}_{1j}$ to generate Y_{1j} from Bernoulli distribution, $B(1, \bar{\bar{p}}_{1j})$. $Y_{1j} = 1$ if the second trial is conducted and zero otherwise.

3. Compute the unconditional and conditional combined effects at stage 2 by

$$\begin{aligned} \hat{\theta}^j_{(2)} =& (\omega_1\hat{\theta}_{1j} + \omega_2 Y_{1j}\hat{\theta}_{2j})/(\omega_1 + \omega_2 Y_{1j}) \\ &\text{and} \\ \bar{\theta}^j_{(2)} =& Y_{1j}(\omega_1\hat{\theta}_{1j} + \omega_2\hat{\theta}_{2j})/(\omega_1 + \omega_2), \end{aligned} \tag{6.12}$$

respectively.

4. Generate the effect size estimate for the third study, $\hat{\theta}_{3j}$ and let ω_3 be its normalised weight so that $\omega_1 = \omega_2 = \omega_3 = 1/3$.

5. Compute $\bar{\bar{p}}_{2j} = \Pr(\hat{\theta}^j_{(2)}, S^2_{1j}, \theta_0)$ the probability of conducting a third trial and generate Y_{2j} from Bernoulli distribution, $B(1, \bar{\bar{p}}_{2j})$. $Y_{2j} = 1$ if the third trial is conducted and zero otherwise.

6. Compute the unconditional and conditional combined effects at stage 3 by

$$\begin{aligned} \hat{\theta}^j_{(3)} =& (\omega_1\hat{\theta}_{1j} + \omega_2 Y_{1j}\hat{\theta}_{2j} + \omega_3 Y_{2j}\hat{\theta}_{3j})/(\omega_1 + \omega_2 Y_{1j} + \omega_3 Y_{2j}) \\ &\text{and} \\ \bar{\theta}^j_{(2)} =& Y_{2j}Y_{1j}(\omega_1\hat{\theta}_{1j} + \omega_2\hat{\theta}_{2j} + \omega_3\hat{\theta}_{3j})/(\omega_1 + \omega_2 + \omega_3), \end{aligned} \tag{6.13}$$

respectively.

7. Repeat steps 1 to 6 for j=1, ..., K, where K is the number of simulations and

calculate the unconditional and conditional means at stages i=2 and 3 by

$$\theta_{(i)} = \sum_{j=1}^{K} \hat{\theta}^{j}_{(i)}/K \text{ and } \mathrm{E}_i = \sum_{j=1}^{K} \bar{\theta}^{j}_{(i)}, \text{ respectively.} \tag{6.14}$$

The entire simulations were conducted based on the following combinations of parameter values; $K = 10000$, θ=(0.00, 0.05, 0.10, ..., 0.95), τ^2=(0.00, 0.02, 0.04, 0.06) and $\theta_0 = 0.5$, $n = (62, 500)$ and the variance σ^2 was taken equal to the sample variance $\hat{\sigma}_1^2 = 19.99$ of the first trial of the example in Section 6.3.

Figures 6.3-6.5 show the biases of the unconditional and conditional expectation of $\hat{\theta}_{(i)}$ for i=1, 2 based on the power-law, extreme value and probit models. For the power-law model, the most biased is the conditional estimator when a new trial is conducted. It can be seen that in this model the unconditional means are reasonably precise, but the step 2 conditional means has a considerable positive bias when the actual effect is small in comparison to the target value θ_0. The step 3 conditional means appear to be even more biased for small values of the actual effect. Increase in τ^2 causes increased bias in the power-law model. Biases for $\sigma^2 = 0.04$ (corresponding to sample size of 500) are given in Figures 6.4 and 6.5. Here the impact of τ^2 is more visible.

For the extreme value and probit models, the unconditional expected values overestimate the mean, and conditional expected values underestimate the actual mean, also bias increases with each step i in decision-making. The biases also increase with increase in τ^2.

Based on the simulations, it can be concluded that different rules and different parameters could give quite different results but these indicate that biases do occur when data dependent rules are used to determine if the next trial should be conducted.

6.1.2.4 The power calculation model for $\Pr(\hat{\theta}, S^2, \theta_0)$

This Section explores the situation where the probability of conducting a new trial depends on power calculations. Consider the case where two studies may be conducted and normally distributed means are used to estimate an effect size θ that is the same in each trial. Typically, if the power calculation yields a small sample size for the second trial, the increase in total power of subsequent meta-analysis will be minor and it may be decided that it is not worth proceeding with the study. Alternatively, the power calculation may yield a large sample size and it may not be possible to achieve the desired power based on the available resources. Let the first study result in an estimate $\hat{\theta}_1$ of θ. The objective is to determine a sample size n_2 for a next study, so that the combined effect $\hat{\theta}_{(2)} = \left(w_1\hat{\theta}_1 + w_2\hat{\theta}_2\right) / (w_1 + w_2)$ will be significantly different from zero (2-sided) at significance level α with $1 - \beta$ power at the target value θ_0. Here the w_i, i=1,2 are the unnormalized weights. The level α should be chosen to account for the multiple testing, but the details of such adjustment are beyond the scope of this work. The variance of the combined effect is then $(w_1 + w_2)^{-1}$. The sample size calculation for the Wald test of the combined effect in meta-analysis is based on setting

$$\frac{\theta_0}{\sqrt{\operatorname{var}\left(\hat{\theta}_{(2)}\right)}} = \theta_0\sqrt{w_1 + w_2} = Z_{1-\alpha/2} + Z_{1-\beta}. \tag{6.15}$$

Now, $w_2 = n_2/\sigma_2^2$ and the resulting equation solved for n_2 yields

$$n_2 = \left(\frac{C^2}{\theta_0^2} - w_1\right)\sigma_2^2, \tag{6.16}$$

where $C = Z_{1-\alpha/2} + Z_{1-\beta}$. The difficulties arise when n_2 is computed using the effect estimate $\hat{\theta}_1$ from the first trial as the effect size θ_0. In order to solve this problem the variance σ_2^2 may be estimated by the sample variance of the first trial, S_1^2. Then the sample size is taken to be

$$n_2 = \frac{C^2 S_1^2}{\hat{\theta}_1^2} - n_1. \tag{6.17}$$

If $\hat{\theta}_1$ is normally distributed and independent of the sample variance S_1^2 which has d_1 degrees of freedom ($d_1 = n_1 - 1$ for one sample, but this notation is introduced for more generality), then $d_1 S_1^2/\sigma_1^2 \sim \chi^2(d_1)$ and the probabilities associated with the experiment can be computed analytically. For example, suppose that it is decided to conduct a new study of size b if $a < n_2 \leq b$ for some points a and b. Then given $\hat{\theta}_1 = \theta_1$, the conditional probability a new trial is conducted is given by

$$\begin{aligned}
\Pr(a, b, \theta_1) &= \Pr\left\{a \leq n_2 \leq b | \hat{\theta}_1 = \theta_1\right\} \\
&= \Pr\left\{a \leq \frac{C^2 S_1^2}{\hat{\theta}_1^2} - n_1 \leq b \middle| \hat{\theta}_1 = \theta_{1.}\right\} \\
&= \Pr\left\{\frac{(a+n_1)\theta_1^2}{C^2} \leq S_1^2 \leq \frac{(b+n_1)\theta_1^2}{C^2}\right\} \\
&= \Pr\left\{\frac{(a+n_1)d_1\theta_1^2}{C^2} \leq \chi^2(d_1) \leq \frac{(b+n_1)d_1\theta_1^2}{C^2}\right\}.
\end{aligned} \tag{6.18}$$

The unconditional probability of conducting a new trial as a function of θ may be computed by integrating the conditional probability over the density

$$f(\theta_1, \theta, \sigma_1^2) \text{ of } \theta_1\text{: } \Pr(a, b) = \int_{-\infty}^{\infty} \Pr(a, b, \theta_1) f(\theta_1 : \theta, \sigma_1^2) d\theta_1. \tag{6.19}$$

If $\hat{\theta}_1 \sim N(\theta, \sigma_1^2/n_1)$ is unbiased, there is no need to perform this integration. Instead, note that $F = n_1\hat{\theta}_1^2/S_1^2$ has a non-central $F_{1,d_1}(\lambda)$ distribution with non-centrality parameter $\lambda = n_1\theta^2/\sigma_1^2$. If a new trial is conducted when $a \leq n_2 \leq b$, the unconditional probability of this is given by

$$\begin{aligned}
\Pr\{a < n_2 \leq b\} &= \Pr\left\{ a \leq \frac{C^2 S_1^2}{\hat{\theta}_1^2} - n_1 \leq b \middle| \hat{\theta}_1 = \theta_1 \right\} \\
&= \Pr\left\{ \frac{(a+n_1)\theta_1^2}{C^2} \leq S_1^2 \leq \frac{(b+n_1)\theta_1^2}{C^2} \right\} \\
&= \Pr\left\{ \frac{(a+n_1)\theta_1^2}{C^2} \leq \frac{n_1\theta_1^2}{F_{1,d_1}(\lambda)} \leq \frac{(b+n_1)\theta_1^2}{C^2} \right\} \\
&= \Pr\left\{ \frac{n_1 C^2}{b+n_1} \leq F_{1,d_1}(\lambda) \leq \frac{n_1 C^2}{a+n_1} \right\}.
\end{aligned} \tag{6.20}$$

The plot of the unconditional and conditional probabilities for $n_1 = 60$, $\sigma_1^2 = 20$, parameters taken from the first trial in the example discussed in Section 6.3 is given in Figure 6.6. It is assumed that $a = 30$ and $b = 80$ in the plots. Figure 6.7 is the plot of the estimated percentage unconditional bias as a function of θ from 10000 simulated experiments for the same scenario with $n_1 = 60$, $\sigma_1^2 = 19.94$ if a second experiment of size 80 is conducted when $30 \leq n_2 \leq 80$. The conditional bias is calculated over the simulations where a second trial was/was not conducted.

Remark 6.1.2. *If the decision to run trial $i+1$ of size b is taken when $a < n_{i+1} \leq b$ for some $i \geq 1$, denote the cumulative combined effect from the first i trials $\hat{\theta}_{(i)} = \sum_{j=1}^{i} w_j\hat{\theta}_j/W_{(i)}$ for $W_{(i)} = \sum_{j=1}^{i} w_j$, and the cumulative sample size $n_{(i)} = \sum_{j=1}^{i} n_j$. Equation* (6.16) *changes to*

$$n_{i+1} = \left(\frac{C^2}{\theta_0} - W_{(i)}\right)\sigma_{i+1}^2, \tag{6.21}$$

where σ_{i+1}^2 is the variance of the trial (i+1). If the variances across trials are assumed equal, this variance can be estimated by the pooled sample variance $S_{(i)}^2$. Also, for reasonably large study sizes, $n_{(i)^*} = W_{(i)}S_{(i)}^2 \approx n_{(i)}$. Under the homogeneity of study *variances, the conditional and unconditional probabilities of a new trial being conducted are approximated by equations* (6.18) *and* (6.20) *with substitution of* $n^*_{(i)}$*, degrees of freedom* $d_{(i)} = n_{(i)} - i$ *and the cumulative effect* $\hat{\theta}_{(i)}$ *instead of* n_1*,* d_1 *and* $\hat{\theta}_1$*, respectively.*

6.2 Sequential design bias

Having made an implicit decision prior to designing the second trial, it important to explore how the knowledge of the results of the first trial used in the design of the second trial affects the combined effect. Resultant bias is referred to as design bias. Suppose in continuation of the approach in Section 6.1, n_2 observations were used to conduct the new experiment instead of the b observations. For the ith trial, i=1, 2, ... with n_i observations, the weights are n_i/σ_i^2, and the effect estimate $\hat{\theta} = \bar{X}_i$. The combined effect over two trials is then $\hat{\theta}_{(2)} = \sum_{i=1}^{2} w_i\hat{\theta}_i / \sum_{i=1}^{2} w_i$. Note that $n_2 = 0$ yields $w_2 = 0$ and $\hat{\theta}_{(2)} = \hat{\theta}_1$. In what follows, assume that the estimates of the effect size and variances are independent, which will hold for samples from normal populations and approximately for other situations.

In practice, σ_2^2 is unknown, therefore a guess value is needed to determine the sample size. Denote this by σ_g^2. A common option is to take $\sigma_g^2 = \sigma_1^2$, which is explored in Section 6.2.1. Then take

$$n_2 = \left(\frac{C^2}{\theta_0} - w_1\right)\sigma_g^2 = \left(\frac{C^2}{\theta_0} - w_1\right)d^2\sigma_2^2, \tag{6.22}$$

where $d^2 = \sigma_g^2/\sigma_2^2$. This is positive if $C^2 > w_1\theta_0^2$ or $n_1 < C^2\sigma_1^2/\theta_0^2$. Let $n_2 = \max(n_2, 0)$. Set $d = 0$ whenever $n_2 = 0$. Therefore, with $w_2 = n_2/\sigma_2^2 = (C^2/\theta_0 - w_1)d^2$, the combined estimate over the two trials is then

$$\hat{\theta}_{(2)} = \frac{w_1\hat{\theta}_1 + w_2\hat{\theta}_2}{w_1 + w_2} = \hat{\theta}_2 + \frac{w_1(\hat{\theta}_1 - \hat{\theta}_2)}{w_1 + w_2}. \tag{6.23}$$

As long as the value of θ_0 used in the sample size calculation is a constant decided by clinical considerations, the expected value of the cumulative effect given by equation (6.23) is equal to θ, and is unbiased. But in the absence of this clinical knowledge, when designing the second trial it is tempting to use the value $\hat{\theta}_1 + \delta$ for some constant δ for the sample size calculation. In clinical trials this can form the basis of proceeding with a phase III trial. That is,

$$n_2 = \left(\frac{C^2}{(\hat{\theta}_1 + \delta)^2} - w_1 \right) \tag{6.24}$$

and $\hat{w}_2 = (C^2/(\hat{\theta}_1 + \delta)^2 - w_1)d^2$. To examine this situation, for simplicity suppose σ_1^2 is known. Note that if $\hat{\theta}_1$ is large then n_2 given by (6.23) can be negative and in this case further experiment is not conducted. Now,

$$E(\hat{\theta}_1 | n_2 \leq 0) = E\left(\hat{\theta}_1 \middle| \hat{\theta}_1 \geq \sqrt{\frac{c^2}{w_1}} - \delta \right) = \theta + \frac{\sigma_1}{\sqrt{n_1}} \frac{\phi(h)}{1 - \Phi(h)},$$

where $h = \{(\sqrt{c^2/w_1} - \delta) - \theta\}/(\sigma_1/\sqrt{n_1})$. For example, if $\theta = 0.2$, $\sigma_1^2 = 19.94$, $\alpha = 0.05$, $\beta = 0.2$, $\delta = 0.2$ and $n_1 = 60$ then $E(\hat{\theta}_1 | n_2 \leq 0) = 3.0848 >> 0.2 = \theta$. That is, if the trials are stopped after the first experiment because the observed result had the desired power for the observed effect size then a highly biased conditional estimate can be obtained. Fortunately, the bias diminishes with increase in n_1, so that for $n_1 = 200$, $\mathrm{E}(\hat{\theta}_1 | n_2 \leq 0) = 1.6506$, and for $n_1 = 1000$, $\mathrm{E}(\hat{\theta}_1 | n_2 \leq 0) = 0.7003$. In the limit, the bias is zero. Similarly,

$$\mathrm{E}(\hat{\theta}_1|n_2 > 0) = \mathrm{E}\left(\hat{\theta}_1\middle|\hat{\theta}_1 < \sqrt{\frac{c^2}{w_1}} - \delta\right) = \theta - \frac{\sigma_1}{\sqrt{n_1}}\frac{\phi(h)}{\Phi(h)}.$$

Then $\mathrm{E}(\hat{\theta}_{(2)}) = \mathrm{E}(\hat{\theta}_1|n_2 \leq 0)\{1 - \Phi(h)\} + \mathrm{E}(\hat{\theta}_{(2)}|n_2 > 0)\Phi(h)$. Suppose that we guess the variance σ^2 exactly; $\mathrm{var}(\hat{\theta}_1) = 1/w_1$. Then

$$\begin{aligned}\mathrm{E}\left(\hat{\theta}_{(2)}|n_2 > 0\right) =& \mathrm{E}\left(\hat{\theta}_{(2)}|\hat{\theta}_1 < \sqrt{\frac{C^2}{w_1}} - \delta\right) \\ =&\theta + \frac{w_1}{C^2}\mathrm{E}\left\{(\hat{\theta}_1 - \theta)(\hat{\theta} - \delta)^2\middle|\hat{\theta}_1 < \sqrt{\frac{C^2}{w_1}} - \delta\right\} \\ \leq&\theta + \frac{w_1}{C^2}\frac{\mathrm{E}\left\{(\hat{\theta}_1 - \theta)(\hat{\theta}_1 + \delta)^2\right\}}{\Phi(h)} \\ =&\theta + \frac{2(\theta + \delta)}{C^2\Phi(h)}.\end{aligned} \tag{6.25}$$

As a consequence of (6.24), it follows that

$$\mathrm{E}(\hat{\theta}_{(2)}) \leq \theta + \left(\frac{\phi(h)\sigma_1}{\sqrt{n_1}} + \frac{2(\theta + \delta)}{c^2}\right), \tag{6.26}$$

giving an upper bound on the bias. That is, if $\delta > 0$ and $\alpha = 0.05$, $\beta = 0.2$ then $c^2 = 7.85$ and the bias is not greater than 25%. In the general case, for an arbitrary d value,

$$\hat{\theta}_{(2)} = \hat{\theta}_2 + \frac{w_1(\hat{\theta}_1 - \hat{\theta}_2)(\hat{\theta}_1 + \delta)^2}{w_1(\hat{\theta}_1 + \delta)^2(1 - d^2) + d^2c^2}.$$

At $d = 0$ this is just $\hat{\theta}_1$ and is unbiased. However, this is of little practical use for at $d = 0$ we would not conduct the second experiment. Now, for an arbitrary d,

$$\mathrm{E}(\hat{\theta}_{(2)}) = \theta + E\left\{\frac{w_1(\hat{\theta}_1 - \theta)(\hat{\theta}_1 + \delta)^2}{w_1(\hat{\theta}_1 + \delta)^2(1 - d^2) + d^2c^2}\right\}, \tag{6.27}$$

which is analytically intractable and we investigate its behaviour using simulations.

6.2.1 Simulations for the sequential design bias

This Section investigates Sequential design bias empirically based on a small simulation study. Specifically, the objective in the simulation is to examine the bias in sequential design relative to different values of d. The simulation was designed as follows.

1. Simulate a trial with n_1 observations using the normal distribution with mean θ_0 and variance σ_1^2.

2. Let $\hat{\theta}_1$ be the sample mean and $\hat{\sigma}_1^2$ the sample variance. Compute $w_1 = n_1/\hat{\sigma}_1^2$.

3. For a given guess σ_g^2

   (a) Compute

   $$n_2 = \left(\frac{C^2}{(\hat{\theta}_1 + \delta)^2} - w_1 \right) \sigma_g^2,$$

   If $n_2 < 5$, replace it with $n_2 = 5$, for $n_2 \geq 1000$, replace it with $n_2 = 1000$.

   (b) Simulate a second trial with n_2 observations. Let $\hat{\theta}_2$ be the sample mean and σ_2^2 be the sample variance.

   (c) Compute $w_2 = n_2/\hat{\sigma}_2^2$ and

   $$\hat{\theta}_{(2)} = \frac{w_1\hat{\theta}_1 + w_2\hat{\theta}_2}{w_1 + w_2}.$$

To evaluate a specific guess such as $\sigma_g^2 = \hat{\sigma}_1^2$, step 3 is carried out once for each simulation. To evaluate the effect of different guesses, step 3 is repeated for different values. As this required different sized trials in step 3 (b) for different guesses, this step was repeated 100 times for each trial in step 1.

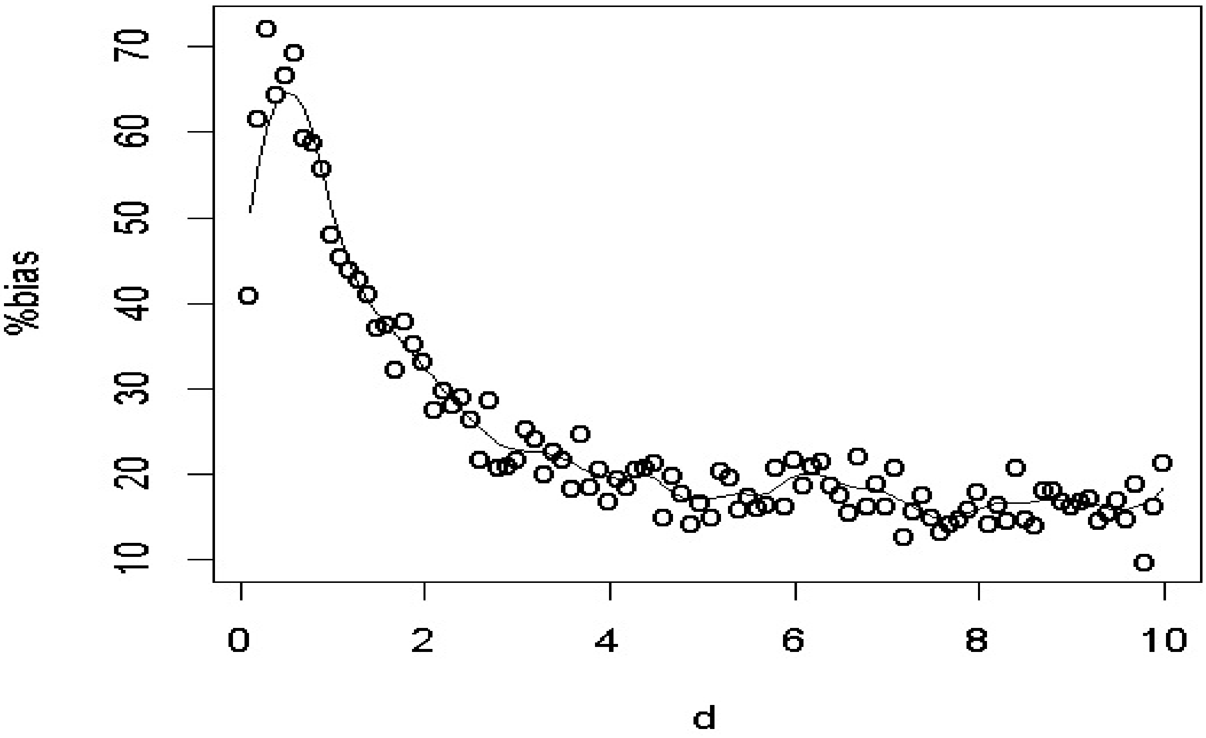


Figure 6.8: Plot of the percent bias against d for simulations described in Section 6.2.1. The parameter values are $\hat{\sigma}_1^2 = 19.94$, $\hat{\sigma}_2^2 = 24.96$, $n_1 = 60$, $\theta = 0.2$, $\delta = 0.2$, $\alpha = 0.05$ and $\beta = 0.2$

The parameter values used for the first simulation are $\theta = 0.2$, $\delta = 0.2$, $\alpha = 0.05$, $\beta = 0.2$, $\sigma_1^2 = 19.94$, $\sigma_2^2 = 24.96$ and $n_1 = 60$ so that $w_1 = 3.00$. Then $c^2/\theta^2 - w_1 = 136.22 > 0$. A total of 1000 simulated initial experiments were conducted, parameter d was taken from 0.1 to 10 in steps of 0.1. 100 second trials were simulated in step 3 (b) for each value of d.

The means of the combined estimators for each value of d are plotted in Figure 6.8.

As there is some variability due to the random sampling, an R (R Core, 2012) package locfit (Loader, 2012) is used to smoothly estimate the mean. It is clear from this plot that the bias can be substantial over a range of guesses for σ_2^2. With $\delta = 0$, the bias was around 15% less but still of concern. If we took $\sigma_g^2 = \hat{\sigma}_1^2$ then the mean percentage bias over the simulations was 49.9% with standard deviation of 3.16 so the bias was uniformly high. As a check, the values of θ and σ_2^2 with $\delta = 0$ were used to compute n_2, and the mean bias was found to be close to zero (0.129%). The bias was 0.042%, also close to zero, when θ and $\hat{\sigma}_1^2$ again with $\delta = 0$ were used to compute n_2, which confirms that the bias arises from using the estimated value of $\hat{\theta}_1$ to decide to carry out the second experiment and compute the sample size.

6.3 Example of Johnson (1993) meta-analysis

This section illustrates the sequential decision bias and sequential design bias using an example of a meta-analysis conducted by Johnson (1993). The meta-analysis comprised 9 studies comparing sodium monouorophosphate (SMFP) to sodium fluoride (NaF) dentifrices in the prevention of caries. The data is referred to as Johnson (1993) meta-analysis. The outcome of interest was the dental score and the effect is estimated by mean difference defined by

$$\hat{\Delta} = \bar{X}_{SMFP} - \bar{X}_{NaF}, \tag{6.28}$$

where $\bar{X}_{SMFP}$ and $\bar{X}_{NaF}$ are the sample means of SMFP and NaF dental scores, respectively. A positive value of $\hat{\Delta}$ favours sodium monouorophosphate (SMFP), and indicates that it is the better of the two dentifrices in the prevention of caries. In order to use equation (6.16), the variances and the sample sizes need to be estimated. Given equal variances $\sigma^2_{1SMFP} = \sigma^2_{1NaF} = \sigma^2_1$ in the treatment (SMFP) and control (NaF) arms of the first trial of sample sizes N_{1SMFP} and N_{1NaF}, respectively, the variance of the difference of the sample means is $\text{var}(\bar{X}_{1SMFP} - \bar{X}_{1NaF}) = \sigma^2_1 \left(N^{-1}_{1SMFP} + N^{-1}_{1NaF}\right) = \sigma^2_1/n_1$ for the effective sample size n_1. The effective sample size is calculated by the geometric mean of the sample sizes of the two arms, $n_1 = \left(N^{-1}_{1SMFP} + N^{-1}_{1NaF}\right)^{-1}$. For the second trial in the two-arms setting, the effective sample size is calculated by the geometric mean of the sample sizes of the two arms N_{2SMFP} and N_{2NaF}. For balanced trial the required sample size is $N_2 = 4n_2$ since $n_2 = (4/N_2)^{-1} = N_2/4$.

Standard fixed-effect meta-analysis of summary data and the cumulative meta-analysis, respectively, obtained with R package ‘meta’ (Schwarzer, 2010) are given in Figure 6.9 (a) and (b). Heterogeneity was not detected (Q=5.38 at 8 degrees of freedom, and $I^2 = 0$), so fixed effect model was used. The first three studies failed to reach significance but showed positive effect. The combined effect after three trials, $\hat{\theta}_{(3)} = 0.52$ is just significant (p-value=0.048), 95% confidence interval [0.01; 1.04]. It will be interesting to see how a decision to continue with the trials could be made during the initial stages of accumulating evidence, and what would be an effect of designing the subsequent trial to show the significantly stronger effect of SMFP.

For the first three trials considered separately, the effective sample sizes were n=(61.30, 81.06, 69.24) and the pooled sample variances S^2=(19.94, 24.96, 8.56), respectively.

Assuming a standard choice of 5% significance level and power of 80%, the constant

$C = z_{\alpha/2} + z_{1-\beta} = 2.802$. After the first trial, the value of n_2 calculated from (6.17) is 150.35. Assume that a new trial of size b is conducted if $a < n_2 < b$ for $a = 50$ and $b = 500$, and the true parameter values are $\theta = 0.28$ (combined effect from 9 trials), and $\sigma^2 = 21.62$ (pooled variance from 9 trials). Under these assumptions, the estimated unconditional probability of continuation of next trial from equation (6.20) is P = 0.349. The F-distribution used to calculate this probability has $d_1 = N - 2$ degrees of freedom, where N is the total sample size of the first trial. If the sample sizes were not restricted from below, take $a = 0$, and the resulting estimated probability of continuation is P = 0.412. To assess the resultant sequential bias, 10000 simulations were performed with $n_1 = 62$ (the effective sample size of the first trial), $\theta = 0.28$ and $\sigma^2 = 19.94$ for $a = 50$ and $b = 500$. From the simulation, the estimated probability of continuation after the first trial is 0.343, percentage unconditional bias after two trials is -19.92%, conditional bias given the decision to stop is -34%, and conditional bias given decision to continue is 8.13%. It is clear from the results that these biases are far from negligible.

If the second trial were run first, the estimated sample size for the next trial is $1718 > b$, and the next trial would not be run. Now suppose that the first two trials were run independently from each other, but the decision is required about the third trial. The variances in these two trials are approximately similar, therefore, the sample size and the probability of continuation with the next trial can be calculated according to Remark 6.1.2. The value of $n^*_{(2)} = 141.27$ is very close to the cumulative effective sample size $n_{(2)} = 142.36$. The sample size calculation from equation (6.21) results in

$n = 376.02$. The conditional probability to continue is 0.25. Taking $a = 0$ increases the probability to 0.30. This probability of continuation is low, and therefore, it is doubtful that the next trial would be funded. But as it happened, the third trial with an effective sample size of 69.24 was run regardless, resulting in marginal significance of the combined effect ($\hat{\theta}_{(3)} = 0.52$, p-value=0.049) of SMFP.

6.4 Discussion of sequential bias in meta-analysis

This Chapter examined sequential bias in accumulating evidence in meta-analysis. Two types of biases were identified, namely sequential decision bias and sequential design bias. It was demonstrated theoretically and by simulations that both sequential decision bias and sequential design bias can arise in sequential and cumulative meta-analyses when the results of previous studies influence the design of a new study. The power-law, extreme value and probit models for determining the probability of conducting the next trial were introduced, followed by a power calculation model. As was demonstrated, biases do occur when data dependent rules are used to determine if the next trial should be conducted. The setting differs from the standard sequential meta-analysis in that a meta-analyst has an active role in the design of the subsequent trial aiming at a definitive meta-analysis. As we have seen, both the conditional and unconditional biases can be non-negligible. Thus caution needs to be exercised in conducting meta-analysis when prior knowledge has been used to design the trials being studied.

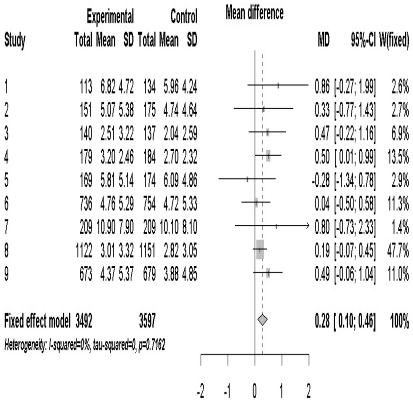
The sample size calculations explored in this Chapter are based on unconditional power of the Wald test for the combined effect. Roloff et al. (2013) advocated using conditional power approach, but the method may not alleviate the bias. The design bias arising from conditional power approach is discussed in context of the designed extension of a clinical trial by Denne (2000). In particular, that paper compared the biases of the estimated effects in conditional and unconditional settings, and found that the differences were minor; see Figure 2 (Denne, 2000).

For designing a new study, a simple fixed effect model is considered for meta-analysis, but even then, the problems of sequential biases become visible. These results should be applicable in random effects model with a necessary change from designing just one study to designing several studies of the same sample size, see Roloff et al. (2013), Kulinskaya and Wood (2014).

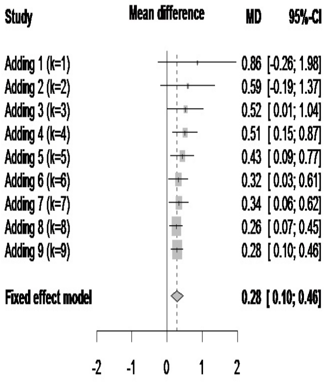
Sequential meta-analysis results in inflation of Type I error due to multiple testing. This problem is not discussed here. A number of procedures aimed at adjustment of significance level to maintain the overall Type I error have been described in Pogue and Yusuf (1997), Wetterslev et al. (2008), Lan and DeMets (1983), van der Tweel and Bollen (2010), Roloff et al. (2013), Higgins et al. (2011). Such adjustment will result in larger sample sizes of the new studies (equations (6.17) and (6.22)) and may decrease sequential biases (equation (6.27)) through increase in critical values at a lower level.

The probability models assumed that a new study is more likely if the existing evidence is in favour of a new treatment than if it is the other way around. However the details of how the interplay between the effect size and the uncertainty may affect the sequential biases were not considered. The value of information (VOI) approach

(Claxton et al., 2002, Claxton and Sculpher, 2006) is an alternative method to decide on the necessity of further research. This method is based on economic modelling comparing the costs involved in further research to benefits of reduction in uncertainty. This method is widely used in contemporary health policy decision- making, (Claxton and Sculpher, 2006). It would be of great interest to investigate the existence of sequential decision bias resulting from this approach.



(a)



(b)

```
Numer of studies combined: k=9


                        MD             95%-CI      z  p.value

Fixed effect model   0.2833  [0.1023; 0.4644] 3.0671   0.0022

b

Quantifying heterogeneity:
```

Chapter 7

Summary and conclusions

Meta-analysis provides accurate estimate of the treatment effect, allows for hypothesis testing and the construction of confidence intervals of the treatment effect. However, temporal changes in magnitude and direction of effect sizes reported in many areas of research (Hodgson et al. (1989), Nieuwkamp et al. (2009), (Hyde et al., 1990), (Gehr et al., 2006), Brugger et al. (2011), Twenge. et al. (2008), Grabe et al. (2008)) can be dramatic and even lead to the loss or gain of the statistical significance of the cumulative treatment effect (Kulinskaya and Koricheva, 2010). Numerous sequential methods have been proposed for monitoring the trends in meta-analysis (Lau et al., 1992, Leimu and Koricheva, 2004, Pogue and Yusuf, 1997, Wetterslev et al., 2008, Higgins et al., 2011, Whitehead, 1997b, Bollen et al., 2006, Kulinskaya and Koricheva, 2010, Lan et al., 2003). However these methods are based on statistical theory applicable only to fixed effect model (FEM) of meta-analysis. For random-effects model (REM), the analysis incorporates the heterogeneity variance, τ^2 and its estimation creates complications.

The main objective in this thesis was to identify an appropriate statistical techniques that are suitable for monitoring trends in random effects model of meta-analysis. This has been achieved by proposing the use of retrospective CUSUM-type test based on sequential procedure by Gombay (2003), Gombay and Serbian (2005) in combination with bootstrap critical values for sequential random-effects meta-analysis. Simulations show that the Type I error rates for the new method are closer to the nominal level in comparison to the existing methods, and are not affected by increase in the level of heterogeneity τ^2.

In random-effects meta-analysis, the heterogeneity of treatment effect across studies creates inferential problems due to non-independence of increments. In the proposed method with bootstrap critical values, the problem does not arise as estimated between-study variance τ^2 is included in the calculation of the bootstrap critical values.

Calculation of bootstrap critical values can be computationally intensive. However, with contemporary high performance computers this should not present much difficulty. Computationally intensive methods involving bootstrapping and permutation tests are becoming common in meta-analysis (Gumedze and Jackson, 2011). An R program for calculating the bootstrap based CUSUM-type test with DerSimonian and Laird (1986), Higgins et al. (2011), Paule and Mandel (1982) and REML estimators of τ^2 is provided in the Appendix.

The drawback of using bootstrap-based critical values is that the resulting method is not true sequential method, and can be used only for retrospective analysis. Even then, it is certainly worthwhile when reviewing the usefulness of an intervention over time. It can be usefully combined with CMA to envisage the trajectory of a cumulative meta-analysis. Unfortunately, as numerous simulations in this work and other authors have repeatedly demonstrated, well-behaved sequential methods for random effects meta-analysis are not yet in existence. In contrast, regardless of the method used to estimate τ^2, the proposed method controls the Type I error irrespective of the number of studies, their sizes and the amount of heterogeneity in treatment effects.

Another issue considered is the effect of accumulating evidence in meta-analysis. Two kinds of bias associated with accumulating evidence, termed "sequential decision bias" and "sequential design bias"were identified. In Chapter 6, it was demonstrated theoretically and by simulation that both sequential decision bias and sequential design bias can arise in sequential and cumulative meta-analysis when the results of previous studies influence the decision to proceed or the design of a new study. Simple models for probability of conducting the next trial were introduced, and an example was provided to demonstrate how decision on whether to continue or stop further trials affects the estimated treatment effect. The setting differs from standard sequential meta-analysis in that a meta-analyst has an active role in the design of the subsequent trial aiming at definitive meta-analysis. Both unconditional and conditional biases were found to be far from negligible. Therefore caution needs to be exercised in conducting meta-analysis when prior knowledge has been used to design new trials being studied.

In clinical trials, favourable results of a phase II trial may be used to design the phase III trial, "*Estimates of treatment effects and variability from earlier trials are traditionally used in the design of trials at the next stage*" (Kirby et al., 2012). This setting is different from meta-analysis in that results are not combined. Moreover, the decision to conduct the phase III trial depends on a significant result in the phase II trial, whereas with meta-analysis guiding research, the sequential trial may be terminated once significance is attained. However, the problem of resulting biases is already recognised in drug development (Wang et al., 2010), and methods of adjustment are sought (Kirby et al., 2012). Perhaps a closer analogy for the sequential decision bias is with group-sequential clinical trials, where a significant result at an interim stage would stop the trial, but otherwise the result of sequential interim stages are accumulated and combined. In this setting the existence of sequential bias is widely recognised and the means of adjustment for this bias have been developed, Whitehead (1986). This adjustment is possible because of the explicit decision rules in these trials. Design bias is similar to the bias induced by mid-trial sample size re-estimation in adaptive trials, Li et al. (2002), Wang et al. (2010). However, methods of sequential bias adjustment in meta-analytic setting are more difficult to develop than in sequential and adaptive clinical trials. The bias depends not only on the unknown true value or the precision of the effect θ, but also on the strategy for making the decision to continue or stop, and of choosing the sample size of the next study. If such a strategy is made explicit, by, say, the Research Council, development of an appropriate bias adjustment should be possible. Development of such a strategy though appears to be an important and complicated problem deserving concerted efforts of statisticians and decision-makers.

To the best of the authors knowledge this is the first time that this important issue is raised in the context of the sequential decision-making associated with the managed accumulation of evidence. Existence of sequential biases raises a number of important research questions. What is the best way to decide on the usefulness of a new trial? How to design this trial so that the resulting combined estimate is the least biased? How to adjust the combined effect to minimise this bias? All these question need to be addressed if evidence-based development of science is to be achieved.

Bibliography

Anscombe, F. (1956). Discussion on Dr. David's and Dr. Johnson's paper. *Journal of the Royal Statistical Society*, 18:24–27.

ARMITAGE, P. et al. (1975). Sequential medical trials. *Sequential medical trials. 2nd edition.*

Armitage, P., McPherson, C., and Rowe, B. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General)*, pages 235–244.

Arnqvist, G. and Wooster, D. (1995). Meta-analysis: synthesizing research findings in ecology and evolution. *Trends in Ecology & Evolution*, 10(6):236–240.

Baker, R. and Jackson, D. (2010). Inference for meta-analysis with a suspected temporal trend. *Biometrical Journal*, 52(4):538–551.

Barnard, G. A. (1959). Control charts and stochastic processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 239–271.

Basu, D. (1977). On the elimination of nuisance parameters. *Journal of the American Statistical Association*, 72(358):355–366.

Becker, L. A. (2000). Effect size. *Accessed on October*, 12(2006):155–159.

Bera, A. and Bilias, Y. (2001). Raos score, Neymans $c(\alpha)$ and silveys LM tests: an essay on historical developments and some new results. *Journal of Statistical Planning and Inference*, 97:9–44.

Biggerstaff, B. and Tweedie, R. (1997). Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in medicine*, 16(7):753–768.

Bollen, C. W., Uiterwaal, C. S., van Vught, A. J., and van der Tweel, I. (2006). Sequential meta-analysis of past clinical trials to determine the use of a new trial. *Epidemiology*, 17(6):644–649.

Brugger, S., Davis, J., Leucht, S., and Stone, J. (2011). Proton magnetic resonance spectroscopy and illness stage in schizophrenia, a systematic review and meta-analysis. *Biological psychiatry*, 69(5):495–503.

Casper, T. C. and Perez, O. A. (2006). An R package for group sequential boundaries using alpha spending functions.

Chalmers, I., Hedges, L. V., and Cooper, H. (2002). A brief history of research synthesis. *Evaluation & the health professions*, 25(1):12–37.

Chow, S.-C., Chang, M., et al. (2008). Adaptive design methods in clinical trials-a review. *Orphanet J Rare Dis*, 3(11):169–90.

Chow, S.-C., Wang, H., and Shao, J. (2007). *Sample size calculations in clinical research*. CRC press.

Claxton, K., Sculpher, M., and Drummond, M. (2002). A rational framework for decision making by the National Institute for Clinical Excellence (nice). *The Lancet*, 360(9334):711–715.

Claxton, K. P. and Sculpher, M. J. (2006). Using value of information analysis to prioritise health research. *Pharmacoeconomics*, 24(11):1055–1068.

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. *Supplement to the Journal of the Royal Statistical Society*, 4(1):102–118.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, 10(1):101–129.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences.

Cooper, H. (2007). Evaluating and Interpreting Research Syntheses in Adult Learning and Literacy. NCSALL Occasional Paper. *National Center for the Study of Adult Learning and Literacy (NCSALL).*

Copas, J. B. (2013). A likelihood-based sensitivity analysis for publication bias in meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(1):47–66.

Corbeil, R. R. and Searle, S. R. (1976). A comparison of variance component estimators. *Biometrics*, pages 779–791.

Denne, J. S. (2000). Estimation following extension of a study on the basis of conditional power. *Journal of biopharmaceutical statistics*, 10(2):131–144.

DerSimonian, R. and Kacker, R. (2007). Random-effects model for meta-analysis of clinical trials: an update. *Contemporary clinical trials*, 28(2):105–114.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–188.

Dobben de Bruyn, C. V. (1968). Cumulative sum tests: theory and practice. *Griffin's Statistical Monographs and Courses*, 24.

Dogo, S. H., Clark, A., and Kulinskaya, E. (2015). A sequential approach for random-effects meta-analysis. *International Journal of Mathematical, Computational, Statistical, Natural and Physical Engineering*, 9(1).

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*, volume 38. SIAM.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American statistical Association*, 82(397):171–185.

Efron, B. and Tibshirani, R. (1985). The bootstrap method for assessing statistical accuracy. Technical report, DTIC Document.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*, volume 57. CRC press.

Field, A. P. (2005). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychological methods*, 10(4):444.

Fisher, R. A. (1934). *Statistical methods for research workers.* Edinburgh.

Gart, J. J., Pettigrew, H. M., and Thomas, D. G. (1985). The effect of bias, variance estimation, skewness and kurtosis of the empirical logit on weighted least squares analyses. *Biometrika*, 72(1):179–190.

Gehr, B., Weiss, C., and Porzsolt, F. (2006). The fading of reported effectiveness. a meta-analysis of randomised controlled trials. *BMC medical research methodology*, 6(1):25.

Gibbons, J. D., Olkin, I., and Sobel, M. (1977). *Selecting and ordering populations.* SIAM.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational researcher*, 5(10):3–8.

Glass, G. V., McGaw, B., and Smith, M. L. (1984). *Meta-analysis in social research.* Sage Newbury Park.

Glasziou, P., Djulbegovic, B., and Burls, A. (2006). Are systematic reviews more cost-effective than randomised trials? *The Lancet*, 367(9528):2057–2058.

Gombay, E. (2003). Sequential change-point detection and estimation. *Sequential Analysis*, 22(3):203–222.

Gombay, E. and Serbian, D. (2005). An adaptation of Pages CUSUM test for change detection. *Periodica Mathematica Hungarica*, 50(1):135–147.

Goudie, A. C., Sutton, A. J., Jones, D. R., and Donald, A. (2010). Empirical assessment suggests that existing evidence could be used more fully in designing randomized controlled trials. *Journal of clinical epidemiology*, 63(9):983–991.

Grabe, S., Ward, L. M., and Hyde, J. S. (2008). The role of the media in body image concerns among women: a meta-analysis of experimental and correlational studies. *Psychological bulletin*, 134(3):460.

Gumedze, F. N. and Jackson, D. (2011). A random effects variance shift model for detecting and accommodating outliers in meta-analysis. *BMC medical research methodology*, 11(1):19.

Hall, P. and Horowitz, J. L. (1996). Bootstrap critical values for tests based on generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, 64(4):891–916.

Hardy, R. J. and Thompson, S. G. (1996). A likelihood approach to meta-analysis with random effects. *Statistics in medicine*, 15(6):619–629.

Harlow, L. L. E., Mulaik, S. A., and Steiger, J. H. (1997). *What if there were no significance tests?* Lawrence Erlbaum Associates Publishers.

Hedges, L. and Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics*, 6(2):107–128.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? the empirical cumulativeness of research. *American Psychologist*, 42(5):443.

Hedges, L. V., Gurevitch, J., and Curtis, P. S. (1999). The meta-analysis of response ratios in experimental ecology. *Ecology*, 80(4):1150–1156.

Hedges, L. V. and Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological methods*, 3(4):486.

Hess, M. R. and Kromrey, J. D. (2004). Robust confidence intervals for effect sizes: A comparative study of Cohens d and Cliffs delta under non-normality and heterogeneous variances. In *Annual Meeting of the American Educational Research Association*.

Higgins, J., Thompson, S. G., Deeks, J. J., and Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, 327(7414):557–560.

Higgins, J., Whitehead, A., and Simmonds, M. (2011). Sequential methods for random-effects meta-analysis. *Statistics in medicine*, 30(9):903–921.

Hoaglin, D. C. (2015). Misunderstandings about Q and Cochran's Q test in meta-analysis. *Statistics in medicine*.

Hodgson, M. J., Parkinson, D. K., and Karpf, M. (1989). Chest X-ray in hypersensitivity pneumonities: A meta analysis of secular trends. *American journal of industrial medicine*, 16(1):45–53.

Horowitz, J. L. (1997). Bootstrap methods in econometrics: theory and numerical performance. *Econometric Society Monographs*, 28:188–222.

Horowitz, J. L. (2001). Handbook of econometrics. *The Bootstrap, Elsevier*, 5:3159–3228.

Horowitz, J. L. and Savin, N. (2000). Empirically relevant critical values for hypothesis tests: A bootstrap approach. *Journal of Econometrics*, 95(2):375–389.

Hu, M., Cappelleri, J. C., and Lan, K. G. (2007). Applying the law of iterated logarithm to control type I error in cumulative meta-analysis of binary outcomes. *Clinical Trials*, 4(4):329–340.

Hubbard, R. and Bayarri, M. (2003). P values are not error probabilities. *Institute of Statistics and Decision Sciences, Working Paper*, (03-26):27708–0251.

Hunter, J. E. and Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models; implications for cumulative research knowledge. *International Journal of Selection and Assessment*, 8(4):275–292.

Hyde, J., Fennema, E., and Lamon, S. (1990). Gender differences in mathematics performance: a meta-analysis. *Psychological bulletin*, 107(2):139.

Ioannidis, J. and Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: The proteus phenomenon in molecular genetics research and randomized trials. *Journal of clinical epidemiology*, 58(6):543–549.

Jackson, D., Turner, R., Rhodes, K., and Viechtbauer, W. (2014). Methods for calculating confidence and credible intervals for the residual between-study variance in random effects meta-regression models. *BMC medical research methodology*, 14(1):103.

Jennions, M. D. and Møller, A. P. (2002). Relationships fade with time: a meta-analysis of temporal trends in publication in ecology and evolution. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 269(1486):43–48.

Jennison, C. and Turnbull, B. W. (2000). *Group sequential tests with applications to clinical trials*. Chapman & Hall/CRC.

Jennison, C. and Turnbull, B. W. (2005). Meta-analyses and adaptive group sequential designs in the clinical development process. *Journal of biopharmaceutical statistics*, 15(4):537–558.

Johnson, M. F. (1993). Comparative efficacy of NaF and SMFP dentrifies in caries prevention: a meta-analytic overview. *Caries review*, 27:328–336.

Khoshdel, A., Attia, J., and Carney, S. (2006). Basic concepts in meta-analysis: a primer for clinicians. *International journal of clinical practice*, 60(10):1287–1294.

Kirby, S., Burke, J., Chuang-Stein, C., and Sin, C. (2012). Discounting phase 2 results when planning phase 3 clinical trials. *Pharmaceutical statistics*, 11(5):373–385.

Koricheva, J., Gurevitch, J., and Mengersen, K. (2013). *Handbook of meta-analysis in ecology and evolution*. Princeton University Press.

Kulinskaya, E. and Koricheva, J. (2010). Use of quality control charts for detection of outliers and temporal trends in cumulative meta-analysis. *Research Synthesis Methods*, 1:297–307.

Kulinskaya, E. and Morgenthaler, S. (2012). *Modern approach to meta-analysis*. The School of professional Development Programme, Imperial College, London.

Kulinskaya, E. and Wood, J. (2014). Trial sequential methods for meta-analysis. *Research Synthesis Methods*, 5(3):212–220.

Kuppens, S. and Onghena, P. (2012). Sequential meta-analysis to determine the sufficiency of cumulative knowledge: The case of early intensive behavioral intervention for children with autism spectrum disorders. *Research in Autism Spectrum Disorders*, 6(1):168–176.

Lai, T. L. (2001). Sequential analysis: some classical problems and new challenges. *Statistica Sinica*, 11(2):303–350.

Lan, K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3):659–663.

Lan, K. G., Hu, M., and Cappelleri, J. C. (2003). Applying the law of iterated logarithm to cumulative meta-analysis of a continuous endpoint. *Statistica Sinica*, 13(4):1135–1146.

Lau, J., Antman, E., Jimenez-Silva, J., Kupelnick, B., Mosteller, F., and Chalmers, T. (1992). Cumulative meta-analysis of therapeutic trials for myocardial infarction. *New England Journal of Medicine*, 327(4):248–254.

Lauritzen, S. (2004). Lectures notes on sequential probability ratio test. *Available online at www.stats.ox.ac.uk/ steffen/teaching/.*

Ledesma, R. D., Macbeth, G., and Cortada de Kohan, N. (2009). Computing effect size measures with vista-the visual statistics system. *Tutorials in Quantitative Methods for Psychology*, 5(1):25–34.

Lehmann, E. (2001). Elements of large-sample theory.

Leimu, R. and Koricheva, J. (2004). Cumulative meta–analysis: a new tool for detection of temporal trends and publication bias in ecology. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271(1551):1961–1966.

Li, G., Shih, W. J., Xie, T., and Lu, J. (2002). A sample size adjustment procedure for clinical trials based on conditional power. *Biostatistics*, 3(2):277–287.

Li, J., Zhang, Q., Zhang, M., and Egger, M. (2007). Intravenous magnesium for acute myocardial infarction. *The Cochrane database of systematic reviews*, (2):CD002755.

Lin, M., Lucas Jr, H. C., and Shmueli, G. (2013). Research commentary-too big to fail: large samples and the p-value problem. *Information Systems Research*, 24(4):906–917.

Lindsey, B. (1983). Efficiency of the conditional score in a mixture setting. *The Annals of Statistics*, pages 486–497.

Loader, C. (2012). Locfit: Local regression, likelihood and density estimation. *R package version 1.5-8.*

McPherson, K. (1974). Statistics: the problem of examining accumulating data more than once. *The New England journal of medicine*, 290(9):501–502.

Montgomery, D. C. (2000). *Introduction To Statistical Quality Control.* Wiley.

Moustakides, G. V. (1986). Optimal stopping times for detecting changes in distributions. *The Annals of Statistics*, 14(4):1379–1387.

Nieuwkamp, D. J., Setz, L. E., Algra, A., Linn, F. H., de Rooij, N. K., and Rinkel, G. J. (2009). Changes in case fatality of aneurysmal subarachnoid haemorrhage over time, according to age, sex, and region: a meta-analysis. *The Lancet Neurology*, 8(7):635–642.

Nowak, R. (2011). Sequential testing. *Available at nowak.ece.wisc.edu/ece830.*

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, 35(3):549–556.

Page, E. (1954). Continuous inspection schemes. *Biometrika*, 14:100–115.

Paule, R. C. and Mandel, J. (1982). Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards*, 87(5):377–385.

Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *The British Medical Journal*, 2(2288):1243–1246.

Pearson, K. (1933). On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika*, 25:379–410.

Petitti, D. B. (1999). *Meta-analysis, decision analysis, and cost-effectiveness analysis: methods for quantitative synthesis in medicine*. Oxford University Press.

Pettigrew, H. M., Gart, J. J., and Thomas, D. G. (1986). The bias and higher cumulants of the logarithm of a binomial variate. *Biometrika*, 73(2):425–435.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199.

Pogue, J. and Yusuf, S. (1997). Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. *Controlled clinical trials*, 18(6):580–593.

Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics*, 51(4):1315–1324.

R Core, T. (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. In *Mathematical Proceedings* *of the Cambridge Philosophical Society*, volume 44, pages 50–57. Cambridge Univ Press.

Rao, P. S., Kaplan, J., and Cochran, W. G. (1981). Estimators for the one-way random effects model with unequal error variances. *Journal of the American Statistical Association*, 76(373):89–97.

Roloff, V., Higgins, J., and Sutton, A. J. (2013). Planning future studies based on the conditional power of a meta-analysis. *Statistics in medicine*, 32(1):11–24.

Rosenthal, R. (1978). How often are our numbers wrong? *American Psychologist*, 33(11):1005.

Rosenthal, R., Cooper, H., and Hedges, L. (1994). Parametric measures of effect size. *The handbook of research synthesis*, pages 231–244.

Rosenthal, R. and DiMatteo, M. R. (2001). Meta-analysis. *Annu. Rev. Psychol.*, 52:59–82.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American psychologist*, 47(10):1173.

Schwarzer, G. (2010). Meta: Meta-analysis with R, R Package version 11–8. *Vienna: The Comprehensive R Archive Network.*

Scrucca, L. (2004). qcc: An R package for quality control charting and statistical process control. *dim (pistonrings)*, 1(200):3.

Serfling, R. J. (1980). Approximation theorems of mathematical statistics.

Shewhart, W. A. (1931). *Economic control of quality of manufactured product*, volume 509. ASQ Quality Press.

Shi, J. Q. and Copas, J. (2004). Meta-analysis for trend estimation. *Statistics in medicine*, 23(1):3–19.

Shu, L. and Jiang, W. (2006). A Markov chain model for the adaptive CUSUM control chart. *Journal of Quality Technology*, 38(2):135.

Siegmund, D. (1985). *Sequential analysis: tests and confidence intervals.* Springer.

Smith, M. L. and Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American psychologist*, 32(9):752.

Stead, L. F., Perera, R., Bullen, C., Mant, D., and Lancaster, T. (2008). Nicotine replacement therapy for smoking cessation. *Cochrane Database Syst Rev*, 1(1).

Sullivan, G. M. and Feinn, R. (2012). Using effect size-or why the p value is not enough. *Journal of graduate medical education*, 4(3):279–282.

Sutton, A. J., Abrams, K. R., Jones, D. R., Jones, D. R., Sheldon, T. A., and Song, F. (2000). *Methods for meta-analysis in medical research.* J. Wiley.

Tippett, L. H. C. et al. (1931). *The methods of statistics.* London: Williams & Norgate Ltd.

Twenge., J. M., Konrath, S., Foster., J. D., Keith Campbell., W., and Bushman., B. J. (2008). Egos inflating over time: A cross-temporal meta-analysis of the narcissistic personality inventory. *Journal of personality*, 76(4):875–902.

van der Tweel, I. and Bollen, C. (2010). Sequential meta-analysis: an efficient decision-making tool. *Clinical Trials*, 0.

Venables, W. (1975). Calculation of confidence intervals for noncentrality parameters. *Journal of the Royal Statistical Society. Series B (Methodological)*, 37(3):406–412.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, 30(3):261–293.

Viechtbauer, W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in medicine*, 26(1):37–52.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3):1–48.

Vorosbcsuk, A. M. (2010). *Systematic review and meta-analysis of outcomes after coronary revascularization procedures with regard to antiplatelet treatment and bleeding complication*. PhD thesis, University of Pecr, Faculty of Medicine, Heart Centre.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186.

Wald, A. and Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, 19(3):326–339.

Wang, Y., Li, G., and Shih, W. J. (2010). Estimation and confidence intervals for two-stage sample-size-flexible design with lsw likelihood approach. *Statistics in Biosciences*, 2(2):180–190.

Wetterslev, J., Thorlund, K., Brok, J., and 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.

Whitehead, A. (1997a). A prospectively planned cumulative meta-analysis applied to a series of concurrent clinical trials. *Statistics in medicine*, 16(24):2901–2913.

Whitehead, J. (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika*, 73(3):573–581.

Whitehead, J. (1997b). *The design and analysis of sequential clinical trials*. John Wiley & Sons.

Whitehead, J. and Stratton, I. (1983). Group sequential clinical trials with triangular continuation regions. *Biometrics*, 39(1):227–236.

Wieringa, J. E. et al. (1999). *Statistical process control for serially correlated data*. PhD thesis, University of Groningen.

Woodall, W. H. (1983). The distribution of the run length of one-sided CUSUM procedures for continuous random variables. *Technometrics*, 25(3):295–301.

Woodward, R. H. and Goldsmith, P. L. (1967). *Cumulative sum techniques*. Oliver & Boyd Edimburgh-London.

Yates, F. and Cochran, W. (1938). The analysis of groups of experiments. *The Journal of Agricultural Science*, 28(04):556–580.

Chapter 8

Appendix

8.1 R programs for SMA, penalised Z-test and retrospective Gombay test

8.1.1 Program for sequential meta-analysis (SMA)

Usage

```
SMA(xT,vT,nT,xC,vC,nC,u0,type.E,level,ty,type.tau,title,stat)
```

Arguments

xT: Number of events in treatment group when data is binary and sample mean when data is continuous.

xC: Number of events in control group when data is binary and sample mean when data is continuous.

vT: Sample variance of treatment group when data is continuous.

vC: Sample variance of control group when data is continuous.

nT: Sample size of treatment group.

nC: Sample size of control group.

type.E: A character string ”MD”, ”SMD”, ”OR”, ”RR” and ”RD” specify the type of effect size measure used.

- ”MD”: Mean difference
- ”SMD”: Standardized mean difference
- ”OR”: Odds ratio
- ”RR”: Relative risk
- ”RD”: Risk difference

ty: A character string ”lt”,”ut” and ”dt” specifying the type of test.

- ”lt”: Lower-sided test
- ”ut”: Upper-sided test
- ”dt”: Double-sided test

type.tau: A character string ”H”, ”DL”, ”MP” and ”REML” specifying the estimator of τ^2 used.

- ”H”: Higgins method
- ”DL”: DerSimonian-Laird estimator
- ”PM”: Paule-Mandel estimator
- ”REML”: Restricted maximum-likelihood estimator

level: Significance level.

u0: Target value.

title: Title of the graph ” .... ”.

stat: Logical statement T or F to provide more statistics below the graph.

Details

This program works in conjunction with the R package ldbounds to calculate the Wald’s test statistic of the cumulative meta-analysis and the group sequential boundaries. The program also provides graphical representation of the results.

Required function

R script below is the function required for the calculations.

```
SMA<-function(xT,vT,nT,xC,vC,nC,type.E,level,u0,ty,type.tau,title,stat){
## This statistics calculate the effect size estimate.
if(type.E=="MD"){
y<-xT-xC
v<-vT/nT+vC/nC
}else{
if(type.E=="SMD"){
N<-nT+nC
pooledva<r-((nT-1)*vT+(nC-1)*vC)/(N-2)
J<-gamma((N-2)/2)/(sqrt((N-2)/2)*gamma((N-3)/2))
y<-J*(xT-xC)/sqrt(pooledvar)
v<-((N-2)*N*J)/((N-4)*nC*nT)+((N-2)*J^2/(N-4)-1)*y^2
}else{
if(type.E=="OR"){
pT<-(xT+.5)/(nT+.5)
pC<-(xC+.5)/(nC+.5)
y<-log(pT*(1-pC)/(pC*(1-pT)))
v<-1/(xT+.5)+1/(nT-xT)+1/(xC+.5)+1/(nC-xC)
}else{
if(type.E=="RR"){
pT<-(xT+.5)/(nT+.5)
pC<-(xC+.5)/(nC+.5)
y<-log(pT)-log(pC)
v<-1/(xT+.5)-1/(nT+.5)+1/(xC+.5)-1/(nC+.5)
}else{
if(type.E=="RD"){
pT<-(xT+.5)/(nT+.5)
pC<-(xC+.5)/(nC+.5)
y<-pT-pC
v<-(xT+.5)*(nT-xT)/(nT+.5)^3+(xC+.5)*(nC-xC)/(nC+.5)^3
}}}}}
K<-length(y)
## This statistic calculates tausquared estimates based on the 4 estimators H,
## DL, MP and REML.
tau<-function(y,v){
K<-length(y)
if(K<2){
H<-0
DL<-0
MP<-0
RM<-0
}else{
tau0<-0.01
eta<-1.5
W<-1/v
```

```
theta<-sum(W*y)/sum(W)
Q<-sum(W*(y-theta)^2)
C<-sum(W)-sum(W^2)/sum(W)
DL<-pmax((Q-(K-1))/(C),0)
H<-(2*(eta-1)*tau0+(K*DL))/(2*(eta-1)+K)
lb<-0
ub<-1000000000000000
f<-function(t,y,v){
sum((1/(v+t))*(y-sum(y/(v+t))/sum(1/(v+t)))^2)-K+1
}
if(f(lb,y=y,v=v)*f(ub,y=y,v=v)<0){
MP<-as.numeric(uniroot(f,c(lb,ub),tol=0.0001,y=y,v=v)[1])
}
else{
MP<-0
}
tau<-0
M<-0
while(M<K){
tauold<-tau
W<-1/(v+tauold)
theta<-sum(W*y)/sum(W)
tau<-pmax((sum((W^2*(((y-theta)^2)-v)))/sum(W^2))+1/sum(W),0)
M<-M+1
}
RM<-tau
}
tau<-c(H,DL,MP,RM)
tau
}
```

```
## This statistic calulate SMA.
n<-(1/nT+1/nC)^(-1)        ## this statistic calculates the effective sample size
K<-length(y)
w<-1/v
theta<-sum(w*y)/sum(w)
Q<-sum(w*(y-theta)^2)
I<-(Q-(K-1))/Q             ## this statistic calculates the rate of inconsistency
Imax<-sum(n)/(1-I)         ## this statistic calculates the heterogeneity
                           ## adjusted optimum information size.
est<-IF<-zval<-numeric(K)
for(k in 1:K){
yk<-y[1:k]
vk<-v[1:k]
nk<-n[1:k]
IF[k]<-sum(nk)/Imax      ## this statistic calculates the information fraction
                         ## at the kth study.
if(type.tau=="H"){
tauhat<-tau(yk,vk)[1]
}else{
if(type.tau=="DL"){
tauhat<-tau(yk,vk)[2]
}else{
if(type.tau=="PM"){
tauhat<-tau(yk,vk)[3]
}else{
if(type.tau=="REML"){
tauhat<-tau(yk,vk)[4]
}}}}
wk<-1/(vk+tauhat)
est[k]<-sum(wk*(yk-u0))/sum(wk)
```

```
zval[k]<-est[k]/sqrt((sum(wk))^(-1))
}
t<-IF
if((ty=="lt")|(ty=="ut")){
level<-2*level
}else{
level<-level
}
## This statistic calculates the Pocock’s boundary values from the R package
## ldbounds.
lb<-round(as.numeric(bounds(t=t,iuse=c(2,2),
    alpha=c(level/2,level/2))$lower.bound),digits=2)
ub<-round(as.numeric(bounds(t=t,iuse=c(2,2),
    alpha=c(level/2,level/2))$upper.bound),digits=2)
if(ty=="lt"){
res<-data.frame(estimates=est,z.value=zval,boundary=lb)
}else{
if(ty=="ut"){
res<-data.frame(estimates=est,z.value=zval,boundary=ub)
}else{
res<-data.frame(estimates=est,z.value=zval,lower.bound=lb,upper.bound=ub)
}}
## These statistics are for the graphical representation
X<-c(1:K)
if((ty=="lt")|(ty=="ut")){
C<-res$boundary
plot(X,zval,axes=FALSE,pch=1,col="black",type="b",ylim=c(min(c(C+.05,zval)),
max(c(C+.05,zval))),xlab=expression(Studies),ylab=expression(paste("zval")),
main=title)
par(new=TRUE)
```

```
plot(X,C,axes=FALSE,pch=1,lty=5,col="red",type="l",ylim=c(min(c(C+.05,zval)),
max(c(C+.05,zval))),xlab=expression(Studies),ylab=expression(paste("zval")),
main=title)
axis(1,at=seq(X[1],max(X),1),labels=TRUE)
axis(2,,labels=TRUE)
box(which="plot",col="blue")
mmin<-min(c(C+.05,zval))-1
if(stat==TRUE){
if(ty=="lt"){
xpoints<-which(zval<=C)
}else{if(ty=="ut"){
xpoints<-which(zval>=C)
}}
if(length(xpoints>0)){
par(mar=c(11,2,2,2))
mtext("Number of times H_0 is rejected: ",side=1,line=6,col="blue",adj=0)
mtext(length(xpoints),side=1,line=7,col="blue",adj=0)
mtext("The first time H_0 is rejected and value of its test statistic:
",side=1,line=9,col="blue",adj=0)
mtext(xpoints[1],side=1,line=10,col="blue",at=2)
mtext(round(zval[xpoints[1]],digits=4),side=1,line=10,col="blue",at=8)
}else{
par(mar=c(11,2,2,2))
mtext("Number of times H_0 is rejected: ",side=1,line=6,col="blue",adj=0)
mtext(length(xpoints),side=1,line=7,col="blue",adj=0)
mtext("The first time H_0 is rejected and value of its test statistic:
",side=1,line=9,col="blue",adj=0)
mtext("NULL",side=1,line=10,col="blue",at=2)
mtext("NULL",side=1,line=10,col="blue",at=8)
}
```

```
}
}else{
C1<-res$lower.bound
C2<-res$upper.bound
plot(X,zval,axes=FALSE,pch=1,col="black",type="b",ylim=c(min(c(C1-.05,zval)),
max(c(C2+.05,zval))),xlab=expression(Studies),ylab=expression(paste("zval")),
main=title)
par(new=TRUE)
plot(X,C1,axes=FALSE,pch=1,col="red",lty=5,type="l",ylim=c(min(c(C1-.05,zval)),
max(c(C2+.05,zval))),
xlab=expression(Studies),ylab=expression(paste("zval")),main=title)
par(new=TRUE)
plot(X,C2,axes=FALSE,pch=1,col="red",lty=5,type="l",ylim=c(min(c(C1-.05,zval)),
max(c(C2+.05,zval))),
xlab=expression(Studies),ylab=expression(paste("zval")),main=title)
axis(1,at=seq(X[1],max(X),1),labels=TRUE)
axis(2,,labels=TRUE)
box(which="plot",col="blue")
if(stat==T){
xpoints<-c(which(zval<=C1),which(zval>=C2))[order(c(which(zval<=C1),
which(zval>=C2)),decreasing=FALSE)]
if(length(xpoints>0)){
par(mar=c(11,2,2,2))
mtext("Number of times H_0 is rejected: ",side=1,line=6,col="blue",adj=0)
mtext(length(xpoints),side=1,line=7,col="blue",adj=0)
mtext("The first time H_0 is rejected and value of its test statistic:
",side=1,line=9,col="blue",adj=0)
mtext(xpoints[1],side=1,line=10,col="blue",at=2)
mtext(round(zval[xpoints[1]],digits=4),side=1,line=10,col="blue",at=8)
}else{
```

```
par(mar=c(11,2,2,2))
mtext("Number of times H_0 is rejected: ",side=1,line=6,col="blue",adj=0)
mtext(length(xpoints),side=1,line=7,col="blue",adj=0)
mtext("The first time H_0 is rejected and value of its test statistic:
",side=1,line=9,col="blue",adj=0)
mtext("NULL",side=1,line=10,col="blue",at=2)
mtext("NULL",side=1,line=10,col="blue",at=8)
}
}
}
}
```

References

DerSimonian, R. and Laired, N. (1986). Meta-analysis in clinical trials. Controlled clinical trials, 7(3):177-188.

Higgins, J., Whitehead, A., and Simmonds, M. (2011). Sequential methods for random-effects meta-analysis. Statistics in medicine, 30(9):903-921.

Paule, R. C. and Mandel, J. (1982). Consensus values and weighting factors. Journal of Research of the National Bureau of Standards, 87(5):377-385.

Wetterslev, J., Thorlund, K.,Brok, J., and 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.

Examples

```
# Example with manually read input of binary data with odd ratio as the efect
```

```
measure
# Source(F:SMA)
# xT<-c(12,9,18,16,45)
# nT<-c(41,50,36,40,59)
# xC<-c(15,7,23,42,34)
# nC<-c(44,39,51,61,77)
# SMA(xT,nT,xC,nC,u0,"OR",level,"lt","DL",title="SMA of REM",stat=T)

# Example with manually read input of continuous data with standardized mean
 difference
# as the efect measure
# Source(F:SMA)
# xT<-c(0.133,0.125,0.087,0.094,0.109,0.062)
# vT<-c(0.025,0.001,0.009,0.098,0.102,0.003)
# nT<-c(41,50,36,40,59,45)
# xC<-c(0.072,0.067,0.100,0.057,0.059,0.107)
# vC<-c(0.053,0.018,0.010,0.068,0.089,0.011)
# nC<-c(44,39,51,61,77)
# SMA(xT,vT,nT,xC,vC,nC,u0,"SMD",level,"lt","DL",title="SMA of REM",stat=T)
```

8.1.2 Program for calculating the penalised Z-test (PZ)

Usage

```
PZ(xT,vT,nT,xC,vC,nC,u0,level,ty,lamda,title,stat)
```

Arguments

xT: Number of events in treatment group when data is binary and sample mean when data is continuous.

xC: Number of events in control group when data is binary and sample mean when data is continuous.

vT: Sample variance of treatment group when data is continuous.

vC: Sample variance of control group when data is continuous.

nT: Sample size of treatment group.

nC: Sample size of control group.

type.E: A character string ”MD”, ”SMD”, ”OR”, ”RR” and ”RD” specifying the type of effect size measure used.

- ”MD”: Mean difference
- ”SMD”: Standardized mean difference
- ”OR”: Odds ratio
- ”RR”: Relative risk
- ”RD”: Risk difference

ty: A character string ”lt”,”ut” and ”dt” specifying the type of test.

- ”lt”: Lower-sided test
- ”ut”: Upper-sided test
- ”dt”: Double-sided test

level: Significance level.

u0: Target value.

lamda: The adjustment constant for penalised Z-test of CMA

title: Title of the graph ” .... ”.

stat: Logical statement T or F to provide more statistics below the graph.

Details

This program calculates the penalized Z-test of the cumulative meta-analysis and provides graphical representation of the results.

Required function

R script below is the function required for the calculation of the penalised Z-test.

```
PZ<-function(xT,vT,nT,xC,vC,nC,type.E,level,u0,ty,lamda,title,stat){
## This statistics calculate the effect size estimate.
if(type.E=="MD"){
y<-xT-xC
v<-vT/nT+vC/nC
}else{
if(type.E=="SMD"){
N<-nT+nC
pooledva<r-((nT-1)*vT+(nC-1)*vC)/(N-2)
J<-gamma((N-2)/2)/(sqrt((N-2)/2)*gamma((N-3)/2))
y<-J*(xT-cC)/sqrt(pooledvar)
v<-((N-2)*N*J)/((N-4)*nC*nT)+((N-2)*J^2/(N-4)-1)*y^2
}else{
if(type.E=="OR"){
pT<-(xT+.5)/(nT+.5)
pC<-(xC+.5)/(nC+.5)
y<-log(pT*(1-pC)/(pC*(1-pT)))
v<-1/(xT+.5)+1/(nT-xT)+1/(xC+.5)+1/(nC-xC)
}else{
if(type.E=="RR"){
pT<-(xT+.5)/(nT+.5)
pC<-(xC+.5)/(nC+.5)
y<-log(pT)-log(pC)
v<-1/(xT+.5)-1/(nT+.5)+1/(xC+.5)-1/(nC+.5)
}else{
if(type.E=="RD"){
pT<-(xT+.5)/(nT+.5)
pC<-(xC+.5)/(nC+.5)
y<-pT-pC
v<-(xT+.5)*(nT-xT)/(nT+.5)^3+(xC+.5)*(nC-xC)/(nC+.5)^3
}}}}}
K<-length(y)
## This statistic calculates tausquared estimate based DL method.
tauhat<-function(y,v){
K<-length(y)
if(K<2){
DL<-0
}else{
W<-1/v
theta<-sum(W*y)/sum(W)
Q<-sum(W*(y-theta)^2)
C<-sum(W)-sum(W^2)/sum(W)
DL<-pmax((Q-(K-1))/(C),0)
}
DL
}
## This statistic calculates the Penalised Z-test of CMA.
tau<-Z<-P<-Ik<-numeric(K)
for(k in 1:K){
yk<-y[1:k]
vk<-v[1:k]
```

```
tau[k]<-tauhat(yk,vk)
if(tau[k]==0){
tau[k]<-var(vk)
}else{
tau[k]<-tau[k]}
if(k==1){
wk<-1/vk
}else{
wk<-1/(tau[k]+vk)}
Ik<-sum(wk)
Sk<-sum(wk*(yk-u0))
if(Sk<=1){
nn<-1
}else{
nn<-log(log(Ik))}
P[k]<-pmax(nn,1)
Z[k]<-Sk/sqrt(lamda*Ik*P[k])
if((ty=="lt")|(ty=="ut")){
level<-level
}else{
level<-level/2}
lb<-rep(qnorm(level,lower.tail=TRUE),K)
ub<-rep(qnorm(level,lower.tail=FALSE),K)
if(ty=="lt"){
res<-data.frame(z.value=Z,boundary=lb)
}else{
if(ty=="ut"){
res<-data.frame(z.value=Z,boundary=ub)}else{
res<-data.frame(z.value=Z,lower.bound=lb,upper.bound=ub)}}
## These statistics are for the graphical representation
```

```
X<-c(1:K)
if((ty=="lt")|(ty=="ut")){
C<-res$boundary
plot(X,Z,axes=FALSE,pch=1,col="black",type="b",ylim=c(min(c(C+.05,Z)),
max(c(C+.05,Z))),xlab=expression(Studies),ylab=expression(paste("zval")),
main=title)
par(new=TRUE)
plot(X,C,axes=FALSE,pch=1,lty=5,col="red",type="l",ylim=c(min(c(C+.05,Z)),
max(c(C+.05,Z))),xlab=expression(Studies),ylab=expression(paste("zval")),
main=title)
axis(1,at=seq(X[1],max(X),1),labels=TRUE)
axis(2,,labels=TRUE)
box(which="plot",col="blue")
if(stat==TRUE){
if(ty=="lt"){
xpoints<-which(Z<=C)
}else{if(ty=="ut"){
xpoints<-which(Z>=C)
}}
if(length(xpoints>0)){
par(mar=c(11,2,2,2))
mtext("Number of times H_0 is rejected: ",side=1,line=6,col="black",adj=0)
mtext(length(xpoints),side=1,line=7,col="black",adj=0)
mtext("The first time H_0 is rejected and value of its test statistic:
",side=1,line=9,col="black",adj=0)
mtext(xpoints[1],side=1,line=10,col="black",at=2)
mtext(round(Z[xpoints[1]],digits=4),side=1,line=10,col="black",at=8)
}else{
par(mar=c(11,2,2,2))
mtext("Number of times H_0 is rejected: ",side=1,line=6,col="black",adj=0)
mtext(length(xpoints),side=1,line=7,col="black",adj=0)
mtext("The first time H_0 is rejected and value of its test statistic:
",side=1,line=9,col="black",adj=0)
mtext("NULL",side=1,line=10,col="black",at=2)
mtext("NULL",side=1,line=10,col="black",at=8)}}}else{
C1<-res$lower.bound
C2<-res$upper.bound
plot(X,Z,axes=FALSE,pch=1,col="black",type="b",ylim=c(min(c(C1-.05,Z)),
max(c(C2+.05,Z))),xlab=expression(Studies),ylab=expression(paste("zval")),
main=title)
par(new=TRUE)
plot(X,C1,axes=FALSE,pch=1,col="red",lty=5,type="l",ylim=c(min(c(C1-.05,Z)),
max(c(C2+.05,Z))),xlab=expression(Studies),ylab=expression(paste("zval")),
main=title)
par(new=TRUE)
plot(X,C2,axes=FALSE,pch=1,col="red",lty=5,type="l",ylim=c(min(c(C1-.05,Z)),
max(c(C2+.05,Z))),xlab=expression(Studies),ylab=expression(paste("zval")),
main=title)
axis(1,at=seq(X[1],max(X),1),labels=TRUE)
axis(2,,labels=TRUE)
box(which="plot",col="blue")
if(stat==T){
xpoints<-c(which(Z<=C1),which(Z>=C2))[order(c(which(Z<=C1),which(Z>=C2)),
decreasing=FALSE)]
if(length(xpoints>0)){
par(mar=c(11,2,2,2))
mtext("Number of times H_0 is rejected: ",side=1,line=6,col="black",adj=0)
mtext(length(xpoints),side=1,line=7,col="black",adj=0)
mtext("The first time H_0 is rejected and value of its test statistic:
",side=1,line=9,col="black",adj=0)
```

```
mtext(xpoints[1],side=1,line=10,col="black",at=2)
mtext(round(Z[xpoints[1]],digits=4),side=1,line=10,col="black",at=8)
}else{
par(mar=c(11,2,2,2))
mtext("Number of times H_0 is rejected: ",side=1,line=6,col="black",adj=0)
mtext(length(xpoints),side=1,line=7,col="black",adj=0)
mtext("The first time H_0 is rejected and value of its test statistic: ",side=1,l
mtext("NULL",side=1,line=10,col="black",at=2)
mtext("NULL",side=1,line=10,col="black",at=8)}}}}
```

References

DerSimonian, R. and Laired, N. (1986). Meta-analysis in clinical trials. Controlled clinical trials, 7(3):177-188.

Hu, M., Cappelleri, J. C., and Lan, K. G (2OO7). Applying the law of iterated algorithm to control Type I error in cumulative meta-analysis of binary outcomes. Clinical Trials, 4(4):329-340.

Lan, K. G. and DeMets, D. L. (2003). Applying the law of iterated algorithm to cumulative meta-analysis of a continuous endpoints. Statistica Sinica, 13(4):1135-1145.

Examples

```
# Example with manually read input of binary data with odd ratio as the efect
measure
# Source(F:PZ)
# xT<-c(12,9,18,16,45)
# nT<-c(41,50,36,40,59)
# xC<-c(15,7,23,42,34)
# nC<-c(44,39,51,61,77)
# PZ(xT,nT,xC,nC,"OR",level,u0,"lt",lamda,title="PZ of REM",stat=T)
# Example with manually read input of continuous data with standardized mean
difference
# as the efect measure
# Source(F:SMA)
# xT<-c(0.133,0.125,0.087,0.094,0.109,0.062)
# vT<-c(0.025,0.001,0.009,0.098,0.102,0.003)
# nT<-c(41,50,36,40,59,45)
# xC<-c(0.072,0.067,0.100,0.057,0.059,0.107)
# vC<-c(0.053,0.018,0.010,0.068,0.089,0.011)
# nC<-c(44,39,51,61,77)
# PZ(xT,vT,nT,xC,vC,nC,"SMD",level,u0,"lt",lamda,title="PZ of REM",stat=T)
```

8.1.3 Program for the bootstrap test

Usage

```
Bootstraptest(xT,vT,nT,xC,vC,nC,type.E,level,u0,ty,type.tau,title,stat)
```

Arguments

xT: Number of events in treatment group when data is binary and sample mean when data is continuous.

xC: Number of events in control group when data is binary and sample mean when data is continuous.

vT: Sample variance of treatment group when data is continuous.

vC: Sample variance of control group when data is continuous.

nT: Sample size of treatment group.

nC: Sample size of control group.

type.E: A character string ”MD”, ”SMD”, ”OR”, ”RR” and ”RD” specifying the type of effect size measure used.

- "MD": Mean difference
- "SMD": Standardized mean difference
- "OR": Odds ratio
- "RR": Relative risk
- "RD": Risk difference

ty: A character string "lt","ut" and "dt" specifying the type of test.

- "lt": Lower-sided test
- "ut": Upper-sided test
- "dt": Double-sided test

type.tau: A character string "H", "DL", "MP" and "REML" specifying the estimator of τ^2 used.

- "H": Higgins
- "DL": DerSimonian-Laird estimator
- "PM": Paule-Mandel estimator
- "REML": Restricted maximum-likelihood estimator

level: Significance level.

u0: Target value.

title: Title of the graph ” .... ”.

stat: Logical statement T or F to provide more statistics below the graph.

Details

This program calculates the Gombay test with bootstrap based critical values for random-effects meta-analysis and produce graphical representation of the results.

Required function

R script below is the function required for the calculations of the bootstrap test.

```
Bootstraptest<-function(xT,vT,nT,xC,vC,nC,type.E,level,u0,ty,type.tau,title,
stat){
## This statistics calculate the effect size estimate.
if(type.E=="MD"){
y<-xT-xC
v<-vT/nT+vC/nC
}else{
if(type.E=="SMD"){
N<-nT+nC
pooledva<r-((nT-1)*vT+(nC-1)*vC)/(N-2)
J<-gamma((N-2)/2)/(sqrt((N-2)/2)*gamma((N-3)/2))
y<-J*(xT-xC)/sqrt(pooledvar)
v<-((N-2)*N*J)/((N-4)*nC*nT)+((N-2)*J^2/(N-4)-1)*y^2
}else{
if(type.E=="OR"){
pT<-(xT+.5)/(nT+.5)
pC<-(xC+.5)/(nC+.5)
y<-log(pT*(1-pC)/(pC*(1-pT)))
v<-1/(xT+.5)+1/(nT-xT)+1/(xC+.5)+1/(nC-xC)
}else{
```

```
if(type.E=="RR"){
pT<-(xT+.5)/(nT+.5)
pC<-(xC+.5)/(nC+.5)
y<-log(pT)-log(pC)
v<-1/(xT+.5)-1/(nT+.5)+1/(xC+.5)-1/(nC+.5)
}else{
if(type.E=="RD"){
pT<-(xT+.5)/(nT+.5)
pC<-(xC+.5)/(nC+.5)
y<-pT-pC
v<-(xT+.5)*(nT-xT)/(nT+.5)^3+(xC+.5)*(nC-xC)/(nC+.5)^3
}}}}}
K<-length(y)
## This statistic calculates tausquared estimates based on the 4 estimators
"H", "DL", "PM" and "REML".
tau<-function(y,v){
K<-length(y)
if(K<2){
H<-0
DL<-0
MP<-0
RM<-0
}else{
tau0<-0.01
eta<-1.5
W<-1/v
theta<-sum(W*y)/sum(W)
Q<-sum(W*(y-theta)^2)
C<-sum(W)-sum(W^2)/sum(W)
DL<-pmax((Q-(K-1))/(C),0)
```

```
H<-(2*(eta-1)*tau0+(K*DL))/(2*(eta-1)+K)
lb<-0
ub<-1000000000000000
f<-function(t,y,v){
sum((1/(v+t))*(y-sum(y/(v+t))/sum(1/(v+t)))^2)-K+1
}
if(f(lb,y=y,v=v)*f(ub,y=y,v=v)<0){
MP<-as.numeric(uniroot(f,c(lb,ub),tol=0.0001,y=y,v=v)[1])
}
else{
MP<-0
}
tau<-0
M<-0
while(M<K){
tauold<-tau
W<-1/(v+tauold)
theta<-sum(W*y)/sum(W)
tau<-pmax((sum((W^2*(((y-theta)^2)-v)))/sum(W^2))+1/sum(W),0)
M<-M+1
}
RM<-tau
}
tau<-c(H,DL,MP,RM)
tau
}
## This statistic calculates the bootstrap based test.
C<-numeric(1)
G<-numeric(K)
## These statistics calculates tau estimate of tau-squared.
```

```
if(type.tau=="H"){
tauhat<-tau(y,v)[1]
}else{
if(type.tau=="DL"){
tauhat<-tau(y,v)[2]
}else{
if(type.tau=="PM"){
tauhat<-tau(y,v)[3]
}else{
if(type.tau=="REML"){
tauhat<-tau(y,v)[4]
}}}}
## The section determines the boostrap critical values.
GGa<-GGb<-numeric(10000)
for (i in 1:10000){
SS<-TT<-numeric(K)
## These statistics generate the bootstrap effect size estimates and the
 sample variances.
if(type.E=="MD"){
bxT<-rnorm(K,u0,sqrt(tauhat+vT))
bxC<-rnorm(K,0,sqrt(tauhat+vC))
TT<-bxT-bxC
SS<-(vT/((nT-1)))*rchisq(K,nT-1)+(vC/((nC-1)))*rchisq(K,nC-1)
}else{
if(type.E=="SMD"){
noncentralpar<-rt(K,N-2,sqrt(nT*nC/N)*y)
y<-rnorm(K,u0,sqrt(tauhat))
TT<-J*N/noncentralpar
SS<-((N-2)*N*J)/((N-4)*nC*nT)+((N-2)*J^2/(N-4)-1)*TT^2
}else{
```

```
if(type.E=="OR"){
xbar<-rnorm(K,u0,sqrt(tauhat))
pTb<-pC*exp(xbar)/(pC*exp(xbar)+1-pC)
xTb<-rbinom(K,nT,pTb)
xCb<-rbinom(K,nC,pC)
TT<-log((xTb+0.5)/(nT+0.5))-log((xCb+0.5)/(nC+0.5))
SS<-(xTb+0.5)^(-1)-(nT+0.5)^(-1)+(xC+0.5)^(-1)-(nC+0.5)^(-1)
}else{
if(type.E=="RR"){
xbar<-pmin(rnorm(K,u0,sqrt(tauhat)),-log(pC))
pTb<-pC*exp(xbar)
xCb<-rbinom(K,nC,pC)
xTb<-rbinom(K,nT,pTb)
pTb<-(xTb+.5)/(nT+.5)
pCb<-(xCb+.5)/(nC+5)
TT<-log(pTb)-log(pCb)
SS<-1/(xTb+.5)-1/(nT+.5)+1/(xCb+.5)-1/(nC+.5)
}else{
if(type.E=="RD"){
xbar<-pmin(pmax(rnorm(K,u0,sqrt(tauhat)),-pC),1-pC)
pTb<-xbar+pC
xCb<-rbinom(K,nC,pC)
xTb<-rbinom(K,nT,pTb)
TT<-xTb/nT-xCb/nC
a<-SS<-numeric(length(nT))
for (mm in 1:length(nT)){
if((xTb[mm]==0)|(xTb[mm]==nT[mm])|(xCb[mm]==0)|(xCb[mm]==nC[mm])){
a[mm]<-.5
}else{
a[mm]<-0
}
SS[mm]<-(xTb[mm]+a[mm])*(nT[mm]-xTb[mm]+a[mm])/(nT[mm]+2*a[mm])^3+(xCb[mm]
+a[mm])*(nC[mm]-xCb[mm]+a[mm])/(nC[mm]+2*a[mm])^3
}
}}}}}

## These statistics re-calculate tau-squared estimate from the bootstrap data.
if(type.tau=="H"){
that<-tau(TT,SS)[1]
}else{
if(type.tau=="DL"){
that<-tau(TT,SS)[2]
}else{
if(type.tau=="PM"){
that<-tau(TT,SS)[3]
}else{
if(type.tau=="REML"){
that<-tau(TT,SS)
}}}}
g<-numeric(K)
for(k in 1:K){
Tk<-SSk<-W<-numeric(k)
Tk<-TT[1:k]
SSk<-SS[1:k]
W<-1/(SSk+that)
#These statistics calculate the Gombay test statistic from bootstrap data.
g[k]<-round(sum(W*(Tk-u0))/sqrt(K*sum(W)),digits=4)
}
GGa[i]<-min(g[1:K])
GGb[i]<-max(g[1:K])
```

```
}
##This statistics compute the bootstrap critical values based on 4 different
estimators of tau.
if(ty=="lt"){    ##lower one-sided critical values
C<- sort(GGa,decreasing=FALSE)[round(length(GGa)*level)]
}else{
if(ty=="ut"){   ##upper one-sided critical values
C<- sort(GGb,decreasing=FALSE)[round(length(GGb)*(1-level))]
}else{    ##two-sided critical values
C<- c(sort(GGa,decreasing=FALSE)[round(length(GGa)*(level/2))],sort(GGb,
decreasing=FALSE)
[round(length(GGb)*(1-level/2))])
}
}
gg<-numeric(K)
## This statistic calculate Gombay test statistic for REM based on the real data.
for(k in 1:K){
wk<-Tk<-Sk<-numeric(k)
Tk<-y[1:k]
Sk<-v[1:k]
wk<-1/(Sk+tauhat)
gg[k]<-round(sum(wk*(Tk-u0))/sqrt(K*sum(wk)),digits=4)
}
## These statistics are for the graphical representation

X<-c(1:K)
if((ty=="lt")|(ty=="ut")){
C<-c(rep(C,K))
results<-data.frame("G"=gg,"bound.G"=C)
plot(X,gg,axes=FALSE,pch=1,col="black",type="b",ylim=c(min(c(C+.05,gg)),
max(c(C+.05,gg))),xlab=expression(Studies),ylab=expression(paste("Gk")),
main=title)
par(new=TRUE)
plot(X,C,axes=FALSE,pch=1,lty=5,col="red",type="l",ylim=c(min(c(C+.05,gg)),
max(c(C+.05,gg))),xlab=expression(Studies),ylab=expression(paste("Gk")),
main=title)
axis(1,at=seq(X[1],max(X),1),labels=TRUE)
axis(2,,labels=TRUE)
box(which="plot",col="blue")
if(stat==TRUE){
if(ty=="lt"){
xpoints<-which(gg<=C)
}else{if(ty=="ut"){
xpoints<-which(gg>=C)
}}
if(length(xpoints>0)){
par(mar=c(11,2,2,2))
mtext("Number of times H_0 is rejected: ",side=1,line=6,col="black",adj=0)
mtext(length(xpoints),side=1,line=7,col="black",adj=0)
mtext("The first time H_0 is rejected and value of its test statistic:
 ",side=1,line=9,col="black",adj=0)
mtext(xpoints[1],side=1,line=10,col="black",at=2)
mtext(round(gg[xpoints[1]],digits=4),side=1,line=10,col="black",at=8)
}else{
par(mar=c(11,2,2,2))
mtext("Number of times H_0 is rejected: ",side=1,line=6,col="black",adj=0)
mtext(length(xpoints),side=1,line=7,col="black",adj=0)
mtext("The first time H_0 is rejected and value of its test statistic:
 ",side=1,line=9,col="black",adj=0)
mtext("NULL",side=1,line=10,col="black",at=2)
mtext("NULL",side=1,line=10,col="black",at=8)
}
}
}else{
C1<-c(rep(C[1],K))
C2<-c(rep(C[2],K))
results<-data.frame("G"=gg,"lower.bound"=C1,"upper.bound"=C2)
plot(X,gg,axes=FALSE,pch=1,col="black",type="b",ylim=c(min(c(C1-.05,gg)),
max(c(C2+.05,gg))),xlab=expression(Studies),ylab=expression(paste("Gk")),
main=title)
par(new=TRUE)
plot(X,C1,axes=FALSE,pch=1,col="red",lty=5,type="l",ylim=c(min(c(C1-.05,gg)),
max(c(C2+.05,gg))),xlab=expression(Studies),ylab=expression(paste("Gk")),
main=title)
par(new=TRUE)
plot(X,C2,axes=FALSE,pch=1,col="red",lty=5,type="l",ylim=c(min(c(C1-.05,gg)),
max(c(C2+.05,gg))),xlab=expression(Studies),ylab=expression(paste("Gk")),
main=title)
axis(1,at=seq(X[1],max(X),1),labels=TRUE)
axis(2,,labels=TRUE)
box(which="plot",col="blue")
if(stat==T){
xpoints<-c(which(gg<=C1),which(gg>=C2))[order(c(which(gg<=C1),which(gg>=C2)),
decreasing=FALSE)]
if(length(xpoints>0)){
par(mar=c(11,2,2,2))
mtext("Number of times H_0 is rejected: ",side=1,line=6,col="black",adj=0)
mtext(length(xpoints),side=1,line=7,col="black",adj=0)
mtext("The first time H_0 is rejected and value of its test statistic:
",side=1,line=9,col="black",adj=0)
```

```
mtext(xpoints[1],side=1,line=10,col="black",at=2)
mtext(round(gg[xpoints[1]],digits=4),side=1,line=10,col="black",at=8)
}else{
par(mar=c(11,2,2,2))
mtext("Number of times H_0 is rejected: ",side=1,line=6,col="black",adj=0)
mtext(length(xpoints),side=1,line=7,col="black",adj=0)
mtext("The first time H_0 is rejected and value of its test statistic:
",side=1,line=9,col="black",adj=0)
mtext("NULL",side=1,line=10,col="black",at=2)
mtext("NULL",side=1,line=10,col="black",at=8)
}
}
}
}
```

References

DerSimonian, R. and Laired, N. (1986). Meta-analysis in clinical trials. Controlled clinical trials, 7(3):177-188.

Higgins, J., Whitehead, A., and Simmonds, M. (2011). Sequential methods for random-effects meta-analysis. Statistics in medicine, 30(9):903-921.

Paule, R. C. and Mandel, J. (1982). Consensus values and weighting factors. Journal of Research of the National Bureau of Standards, 87(5):377-385.

Gombay, E. (2003). Sequential change-point detection and estimation. Sequential Analysis 22(3):203-222.

Gombay, E. and Serbian, D. (2005). An adaptation of Page CUSUM test for change detection. Periodica Mathematica Hungarica, 50(1):135-147.

Examples

```
# Example with manually read input of binary data with odd ratio as the
 efect measure
# Source(F:bootstraptest)
# xT<-c(12,9,18,16,45)
# nT<-c(41,50,36,40,59)
# xC<-c(15,7,23,42,34)
# nC<-c(44,39,51,61,77)
# Bootstraptest(xT,nT,xC,nC,type.E="OR",level,u0,ty,type.tau,
title="Bootstrap based tes",
# stat=T)

# Example with manually read input of continuous data with standardized
mean difference
# as the efect measure
# Source(F:SMA)
# xT<-c(0.133,0.125,0.087,0.094,0.109,0.062)
# vT<-c(0.025,0.001,0.009,0.098,0.102,0.003)
# nT<-c(41,50,36,40,59,45)
# xC<-c(0.072,0.067,0.100,0.057,0.059,0.107)
# vC<-c(0.053,0.018,0.010,0.068,0.089,0.011)
# nC<-c(44,39,51,61,77)
# Bootstraptest(xT,vT,nT,xC,vC,nC,type.E="OR",level,u0,ty,type.tau,
# title="Bootstrap based test",stat=T)
```

8.2 R programs for calculations in Chapter 6

8.2.1 Program for unconditional and conditional probability presented in Figure 6.6

```
## DESCRIPTION
################
## This program calculates the unconditional and conditional probabilities (Figure 6.1) as a
function of the true theta of theta0 conducting the second trial for the power
calculation rule discussed in Section 6.1.2.5.
#########################################################################################

## USAGE
## F2(theta,a,b,n.1,sigma.1,alpha,beta)

## ARGUREMENT

## theta       A vector specifying the values of theta, the effect parameter.
## sigma.1     Within-study variance of study 1.
## alpha       Value of Type I error.
## beta        Value of Type II error.
## n.1         Sample size of study 1.
## a           A positive integer specifying a lower bound for sample size of study 2.
## b           A positive interger specifying an upper bound for sample size of study 2.

################################################################################################

## PROGRAM

F2<-function(theta,a,b,n.1,sigma.1,alpha,beta){
a.0<-qnorm(alpha/2, lower.tail=FALSE)
b.0<-qnorm(1-beta,lower.tail=TRUE)
c2<-(a.0+b.0)^2
par(mfrow=c(2,1))
pr=function(theta,a,b,n.1,c2,sigma.1)
```

```
    {
        ell=(a+n.1)*theta^2*(n.1-1)/(c2*sigma.1);
        u=(b+n.1)*theta^2*(n.1-1)/(c2*sigma.1);
        pchisq(u,n.1-1)-pchisq(ell,n.1-1)
    }

pr.2=function(th,a,b,n.1,c2,sigma.1,theta)
    {
        pr(th,a,b,n.1,c2,sigma.1)*dnorm(th,theta,sqrt(sigma.1/n.1));
    }
Pr=NULL;
for(th in theta)
    {
        Pr=c(Pr,pr(th,a,b,n.1,c2,sigma.1));
    }
Pr.1=NULL;
for(theta.1 in theta)
    {
        Pr.1=c(Pr.1,integrate(pr.2,lower=-5,upper=5,a=a,b=b,n.1=n.1,c2=c2,sigma.1=sigma.1,
        +theta.1)$value);
    }
plot(theta,Pr,type="l",ylab="Probability",main="Conditional",xlab=expression(theta));
plot(theta,Pr.1,type="l",ylab="Probability",main="unconditional",xlab=expression(theta));
}
```

8.2.2 Program for % bias presented in Figure 6.7

```
##DESCRIPTION
###########################
## This program simulate the percentage bias of the unconditional probability power rule
when d=1 or for Figure 3
###############################################################################################

## USAGE
## F3(n.sim,mu,a,b,n.1,sigma.1,delta,alpha,beta)

## ARGUREMENT
## n.sim       A positive integer specifying the number of simulations to be carried out.
## mu          A vector specifying the values of the effect parameter.
## sigma.1     Within-study variance of study 1.
## a           A positive integer specifying a lower bound for sample size of study 2.
## b           A positive interger specifying an upper bound for sample size of study 2.
## n.1         Sample size of study 1.
## delta       Real number specifying the size of the effect
## alpha       Value of Type I error.
## beta        Value of Type II error.

F3<-function(n.sim,mu,a,b,n.1,sigma.1,delta,alpha,beta){
a.0<-qnorm(alpha/2, lower.tail=FALSE)
b.0<-qnorm(1-beta,lower.tail=TRUE)
c2<-(a.0+b.0)^2

sim.f <- function(theta.0,n.1,n.sim,sigma.1,a,b,delta=0 )
  {
    X <- matrix(rnorm(n.1*n.sim,theta.0,sqrt(sigma.1)),nrow=n.sim);
    theta.1.hat <- apply(X,1,mean);
    sigma.1.hat <- apply(X,1,var);
    mean(theta.1.hat);
    w.1 <- n.1/sigma.1.hat;

    n.2 <- (c2/(theta.1.hat+delta)^2-w.1)*sigma.1.hat;

    t.1 <- (a<=n.2)&(n.2<=b);
    n.t <- sum(t.1);
    if(n.t>0)
      {
        Y <- matrix(rnorm(b*n.t,theta.0,sqrt(sigma.1)),nrow=n.t);
        theta.2.hat <- apply(Y,1,mean);
        sigma.2.hat <- apply(Y,1,var);

        th <- rep(0,n.sim);
        aa <- as.numeric(t.1);

        w.2 <- b/sigma.2.hat;

        theta.c.1 <- (theta.1.hat[t.1]*w.1[t.1]+theta.2.hat*w.2)/(w.1[t.1]+w.2)

        th[t.1] <- theta.c.1;

        theta.c <- theta.1.hat*(1-aa)+th

        u.m <- 100*(mean(theta.c,na.rm=T)-theta.0)/theta.0;
        c.m <- 100*(mean(theta.c[t.1],na.rm=T)-theta.0)/theta.0;
        c.m.0 <- 100*(mean(theta.c[!t.1],na.rm=T)-theta.0)/theta.0;
        pp <- sum(t.1)/length(t.1);
        }else{u.m <- NA;c.m <- NA;c.m.0 <- NA;pp <- NA}
    list(u.m=u.m,c.m=c.m,c.m.0=c.m.0,pp=pp);
  }
#########################

####w.10 <- n.1/sigma.1;

######################

B <- NULL;
B.C.1 <- NULL;
B.C.0 <- NULL;
PP <- NULL;

for(k in 1:length(mu))
  {
    s.out <- sim.f(mu[k],n.1,n.sim,sigma.1,a,b);
    B <- c(B,s.out$u.m);
    B.C.1 <- c(B.C.1,s.out$c.m);
    B.C.0 <- c(B.C.0,s.out$c.m.0);
    PP <- c(PP,s.out$pp)
  }
matplot(mu,cbind(B,B.C.1,B.C.0),type="b",xlab="theta",ylab="% bias",pch=c("*","+","#"),
cex=1,col=c(1,1,1));legend(0.9,70,legend=c("Unconditional","Y=1","Y=0")
,col=c(1,1,1),pch=c("*","+","#"))#lty=c(1,2,3))

matplot(mu,cbind(B,B.C.1,B.C.0),type="l",xlab=expression(theta),ylab="% bias",lwd=2,
lty=c(1,3,4));legend(0.9,50,legend=c("Unconditional","Y=1","Y=0")
,col=c(1,2,3),lty=c(1,3,4),lwd=2);#,pch=c("*","+","#"))#lty=c(1,2,3))
abline(h=0.00, lty=1, lwd=2,col="lightgrey")

##################################################
theta.0 <- 0.3;
h <- (sqrt(c2/w.10)-delta-theta.0)/(sqrt(sigma.1/n.1));
ph <- dnorm(h);
PH <- pnorm(h);
mu.0 <- theta.0+sqrt(1/w.10)*ph/(1-PH);

mu.1 <- theta.0-sqrt(1/w.10)*ph/PH;

c(mu.0,mu.1)

c(mean(theta.1.hat[!t.1]),mean(theta.1.hat[t.1]))

}
```

8.2.3 Program for simulations for sequential design bias in Section 6.2.1 and Figure 6.8

```
## DESCRIPTION
###############
## This program works in conjunction with R package logfit. It calculates and plots
the percentage bias of the simulations discussed in Section 3
###############################################################################################

## USAGE
```

```
F4(n.sim,theta.0,d.all,n.1,sigma.1,sigma.2,delta,alpha,beta)

## ARGUREMENT

## n.sim       A positive integer specifying the number of simulations to be carried out.
## theta.o     The null value of the effect parameter.
d.all          A vector of real numbers specifying the values of d used in the simulations.
## n.1         Sample size of study 1.
## sigma.1     Within-study variance of study 1.
## sigma.2     True value of within-study variance of study 2.
## delta       Real number specifying the size of the effect
## alpha       Value of Type I error.
## beta        Value of Type II error.


##############################
F4<-function(n.sim,theta.0,d.all,n.1,sigma.1,sigma.2,delta,alpha,beta){

a.0<-qnorm(alpha/2, lower.tail=FALSE)
b.0<-qnorm(1-beta,lower.tail=TRUE)
c2<-(a.0+b.0)^2
w.10 <- n.1/sigma.1;

S.1 <- matrix(rnorm(n.sim*n.1,theta.0,sqrt(sigma.1)),ncol=n.1);

theta.1.hat <- apply(S.1,1,mean);
sigma.1.hat <- apply(S.1,1,var);
w.1 <- n.1/sigma.1.hat;

c(theta.0,mean(theta.1.hat))

M <- NULL;
P <- NULL;
for ( d in d.all)
{
```

```
################################
##Study 2
###############################

  sigma.g <- d^2*sigma.2;
  n.2 <- (c2/(theta.1.hat+delta)^2-w.1)*sigma.g;###first simulations
#  n.2 <- (c2/(theta.1.hat+delta)^2-w.1)*sigma.1.hat;
#  n.2 <- (c2/(theta.0)^2-w.1)*sigma.2;
#  n.2 <- (c2/(theta.0)^2-w.1)**sigma.1.hat
#  n.2 <- rep(n.2,n.sim)
  n.2 <- ceiling(n.2);
  n.2 <- pmax(n.2,5);
  n.2 <- pmin(n.2,1000)

  theta.2.hat <- NULL;
  sigma.2.hat <- NULL;

  for(k in 1:n.sim)
    {
      S.2 <- rnorm(n.2[k],theta.0,sqrt(sigma.2))
      theta.2.hat <- c(theta.2.hat,mean(S.2))
      sigma.2.hat <- c(sigma.2.hat,var(S.2))
    }
#################################
##combined estimator and bias
#################################

  w.2 <-  n.2/sigma.2.hat

  theta.c <- (w.1*theta.1.hat+w.2*theta.2.hat)/(w.1+w.2);

  perc.bias <- (mean(theta.c)/theta.0-1)*100
  P <- c(P,perc.bias);
  M <- c(M,mean(theta.c));
}
###################################
```

```
## plot(d.all,M,type="l");
## plot(d.all,P,xlab="d",ylab="% bias",type="l");

library(locfit);

L.P <- locfit(P~lp(d.all,nn=0.15))
plot(L.P,get.data=T,xlab="d",ylab="% bias");

h <- (sqrt(c2/w.10)-delta-theta.0)/sqrt(sigma.1);
t.1 <- dnorm(h)/(1-pnorm(h))
theta.0.t <- theta.0+t.1*sqrt(sigma.1/n.1)

B=data.frame("c2/theta^2-n.1"=c2/theta.0^2-n.1,h=h,t.1=t.1)
B
}
```

8.2.4 Program for the simulations of unconditional and conditional means in Section 6.1.2.3

```
## DESCRIPTION
##############
## This program simulates the unconditional and conditional means of the combined effect
 at stages 2 and 3 discussed in Section 2.
#######################################################################################


## USAGE
## T(nsim,theta,theta0,sigma.1,t,n.1,tau,r).


## ARGUREMENTS
##############

## nsim       A positive number specifying the number of simulations.
## theta      The TRUE value of the effect parameter.
## theta0     The target value of the effect parameter.
## sigma.1    Within-study variance of study 1
## t          A positive integer specifying the power-index of the power-law model.
## n.1        The sample size of study 1
## tau        Between-study variance.
## r          r is a real number to determine
#################################################################################################

## PROGRAM

T<-function(nsim,theta,theta0,sigma.1,t,n.1,tau,r){
# This statistic generates the effect size for study 1
theta1<-rnorm(nsim,theta,sqrt(sigma.1/n.1+tau))
# This statistic calculates the probability of conducting the second trial for power model
p11<-(theta1/theta0)^t
p11<-p11*(1-(p11>1))
p11<-p11*(1-(p11<0))



# This statistic calculates the probability of conducting the second trial for extreme value
model
x<-theta1
GA<-exp(-exp((x-theta0)/sqrt(sigma.1/N+tau)))
GB<-exp(-exp((r*theta0-theta0)/sqrt(sigma.1/N+tau)))
GA[which(x<r*theta0)]<-0
p12<-GA/GB

# This statistic calculates the probability of conducting the second trial for probit model
x<-theta1
alpha<-0
beta<-1
G1<-1-pnorm((x-theta0)/sqrt(sigma.1/N+tau))
G2<-1-pnorm((r*theta0-theta0)/sqrt(sigma.1/N+tau))
G1[which(x<r*theta0)]<-0
p13<-G1/G2

# If Y1i=1 for i=1, 2, 3, the second trial is conducted otherwise it is not conducted.
```

```
Y11<-rbinom(nsim,1, p11)
Y12<-rbinom(nsim,1, p12)
Y13<-rbinom(nsim,1, p13)



theta21<-Y11*rnorm(nsim,theta,sqrt(sigma.1/n.1+tau))
theta22<-Y12*rnorm(nsim,theta,sqrt(sigma.1/n.1+tau))
theta23<-Y13*rnorm(nsim,theta,sqrt(sigma.1/n.1+tau))

theta21cum<-(theta1+theta21)/(1+Y11)
theta22cum<-(theta1+theta22)/(1+Y12)
theta23cum<-(theta1+theta23)/(1+Y13)



theta21cond<-(theta1+theta21)*Y11/2
theta22cond<-(theta1+theta22)*Y12/2
theta23cond<-(theta1+theta23)*Y13/2

# This statistic calculates the probability of conducting the second trial for power model
p21<-(theta21cond/theta0)^t
p21<-p21*(1-(p21>1))
p21<-p21*(1-(p21<0))



# This statistic calculates the probability of conducting the second trial for
 extreme value model
x<-theta22cond
GGA<-exp(-exp((x-theta0)/sqrt(sigma.1/N+tau)))
GGB<-exp(-exp((r*theta0-theta0)/sqrt(sigma.1/N+tau)))
GGA[which(x<r*theta0)]<-0
p22<-GGA/GGB

# This statistic calculates the probability of conducting the second trial for probit model
x<-theta23cond
alpha<-0
beta<-1
```

```
GG1<-1-pnorm((x-theta0)/sqrt(sigma.1/N+tau))
GG2<-1-pnorm((r*theta0-theta0)/sqrt(sigma.1/N+tau))
GG1[which(x<r*theta0)]<-0
p23<-GG1/GG2

Y21<-rbinom(nsim,1, p21)
Y22<-rbinom(nsim,1, p22)
Y23<-rbinom(nsim,1, p23)

theta31<-Y11*Y21*rnorm(nsim,theta,sqrt(sigma.1/n.1+tau))
theta32<-Y12*Y22*rnorm(nsim,theta,sqrt(sigma.1/n.1+tau))
theta33<-Y13*Y23*rnorm(nsim,theta,sqrt(sigma.1/n.1+tau))

theta31cum<-(theta1+theta21+theta31)/(1+Y11+Y11*Y21)
theta32cum<-(theta1+theta22+theta32)/(1+Y12+Y12*Y22)
theta33cum<-(theta1+theta23+theta33)/(1+Y13+Y13*Y23)

theta31cond<-(theta1+theta21+theta31)*Y11*Y21/3
theta32cond<-(theta1+theta22+theta32)*Y12*Y22/3
theta33cond<-(theta1+theta23+theta33)*Y13*Y23/3

# These statististics calculates the unconditional expected value of the combined effect
 at the second trial
e21<-round(mean(theta21cum),digits=4) # based on power-law model
e22<-round(mean(theta22cum),digits=4) # based on extreme value model
e23<-round(mean(theta23cum),digits=4) # based on probit model

# These statististics calculate the conditional expected value of the combined effect
at the second trial
E21<-round(sum(theta21cond)/sum(Y11),digits=4)   # based on power-law model
E22<-round(sum(theta22cond)/sum(Y12),digits=4)   # based on extreme value model
E23<-round(sum(theta23cond)/sum(Y13),digits=4)   #based on probit model
d21<-sum(Y11)  # based on power-law model
d22<-sum(Y12)  # based on extreme value model
d23<-sum(Y13)  # based on probit model
```

```
# These statististics calculate the unconditional expected value of the combined effect
at the third trial
e31<-round(mean(theta31cum),digits=4) # based on power-law model
e32<-round(mean(theta32cum),digits=4) # based on extreme value model
e33<-round(mean(theta33cum),digits=4) # based on probit model

# These statististics calculate the conditional expected value of the combined effect
at the third trial
E31<-round(sum(theta31cond)/sum(Y11*Y21),digits=4)  # based on power-law model
E32<-round(sum(theta32cond)/sum(Y12*Y22),digits=4)  # based on extreme value model
E33<-round(sum(theta33cond)/sum(Y13*Y23),digits=4)  # based on probit model
d31<-sum(Y11*Y21) # based on power-law model
d32<-sum(Y12*Y22) # based on extreme value model
d33<-sum(Y13*Y23) # based on probit model
B<-data.frame(stage=c(2,2,2,3,3,3),model.Type=c("power-law","extreme value","probit",
"power-law","extreme value","probit"),uncond.means=c(e21,e22,e23,e31,e32,e33),
cond.means=c(E21,E22,E23,E31,E32,E32),No.sudies.use=
c(d21,d22,d23,d31,d32,d33))
B
}
```

8.3 Data and results of calculations in Chapter 5

Table 8.1: Data and results of the meta-analysis of 23 studies on magnesium for myocardial infarction by Li et al. (2007). The subscripts T and C refer to the treatment and control arms of the studies. The columns headed n_T and n_C are the sample sizes, and x_T and x_C are the numbers of events in each study. The columns headed φ, v and φ_{cum} are the log-odds ratios, their variances and the cumulative effects, respectively. The next two columns ($z_{-0.934}$ and $B_{-0.934}$) are the results of SMA; Z_c are the values of the Z-test for CMA with target value c and B_c are the Pocock's boundaries for the SMA. The next column ($L_{(-0.934,1.64)}$) is the results of penalized Z-test; $L_{(c,b)}$ are the values of the penalized Z-test with target value c and critical value b. Finally the last eight columns ($GH_{(-0.934,0.61)}$, $GDL_{(-0.934,0.61)}$, $GMP_{(-0.934,0.56)}$, $GREML_{(-0.934,0.54)}$) are the results of Gombay test for REM. $GH_{(c,b)}$, $GDL_{(c,b)}$, $GMP_{(c,b)}$, and $GREML_{(c,b)}$ are the results of Gombay test for REM based on $\hat{\tau^2_H}$, $\hat{\tau^2_{DL}}$, $\hat{\tau^2_{MP}}$ and $\hat{\tau^2_{REML}}$, respectively, c is the target value and b is the bootstrap critical value.

S/N	Author (Year)	xT	nT	xC	nC	$\hat{\varphi}$	v	φ_{cum}	$z_{-0.93}$	$B_{-0.93)}$	$L_{(-0.934,1.6448)}$
1	Morton (1984)	1	40	2	36	-0.65	1.12	-0.65	0.2657	3.97	0.2204
2	Rasmussen (1986)	4	56	14	74	-1.03	0.33	-0.94	-0.0045	3.84	0.0296
3	Smith (1986)	2	92	7	93	-1.14	0.56	-1.01	-0.1576	3.73	-0.1025
4	Abraham (1987)	1	48	1	46	-0.04	1.38	-0.90	0.1090	3.80	0.1296
5	Ceremuzynski (1989)	1	25	3	23	-1.03	1.04	-0.91	0.0679	3.84	0.0856
6	Singh (1990)	6	81	11	81	-0.64	0.27	-0.82	0.3578	3.70	0.2596
7	Shechter (1990)	1	50	9	53	-1.95	0.82	-0.93	-0.0159	3.72	-0.0336
8	Feldsted (1991)	10	150	8	148	0.21	0.23	-0.63	1.0997	3.56	0.6428
9	Shechter (1991)	2	21	4	25	-0.49	0.72	-0.62	1.1996	3.66	0.7552
10	Woods (1992)	90	1150	118	1150	-0.30	0.02	-0.39	2.7237	3.07	1.1797
11	Wu (1992)	5	125	12	102	-1.11	0.28	-0.43	2.5141	3.23	1.0386
12	Bhargava (1995)	3	40	3	38	-0.06	0.63	-0.42	2.6780	3.29	1.2195
13	Shechter (1995)	4	96	17	98	-1.49	0.30	-0.57	2.3274	3.28	1.6448
14	Thogersen (1995)	4	130	8	122	-0.74	0.36	-0.54	2.3349	3.27	1.9990
15	ISIS-4a (1995)	928	11675	880	11648	0.06	0.002	-0.51	4.8769	2.34	1.8207

S/N	Author (Year)	xT	nT	xC	nC	$\hat{\varphi}$	v	φ_{cum}	$z_{-0.93}$	$B_{-0.93)}$	$L_{(-0.934,1.6448)}$
16	ISIS-4b (1995)	1288	17333	1223	17390	0.06	0.002	-0.22	6.7940	2.19	4.8188
17	Urek (1996)	1	31	0	30	1.10	2.73	-0.21	6.8563	2.25	4.9268
18	Raghu (1999)	6	169	18	181	-1.05	0.22	-0.26	6.6768	2.29	4.4140
19	Gyamlani (2000)	2	50	10	50	-1.62	0.54	-0.31	6.5017	2.32	3.9369
20	MAGIC (2000)	475	3113	472	3100	0.003	0.01	-0.20	7.8455	2.33	5.6364
21	Santoro (2000)	0	75	1	75	-1.11	2.69	-0.20	7.8282	2.38	5.7141
22	Zhu (2002)	101	1691	134	1488	-0.44	0.02	-0.26	8.0590	2.38	5.1986
23	Nakashima (2004)	1	89	3	91	-0.85	0.98	-0.26	8.0384	2.48	5.1831

Table 8.2: Data and results of the meta-analysis of 23 studies on magnesium for myocardial infarction by Li et al. (2007). cont'd

S/N	$GH_{(0,-0.49)}$	$GDL_{(0,-0.50)}$	$GMP_{(0,-0.49)}$	$GREML_{(0,-0.49)}$	$GH_{(-0.934,1.01)}$	$GDL_{(-0.934,1.02)}$	$GMP_{(-0.93,0.95)}$	$GREML_{(-0.934,0.}$
1	-0.1256	-0.1255	-0.1226	-0.1199	0.0554	0.0554	0.0541	0.0529
2	-0.3714	-0.3709	-0.3469	-0.3273	-0.0010	-0.0009	0.0016	0.0037
3	-0.4816	-0.4810	-0.4538	-0.4311	-0.0329	-0.0329	-0.0296	-0.0269
4	-0.4550	-0.4544	-0.4271	-0.4043	0.0227	0.0227	0.0295	0.0295
5	-0.4990	-0.4984	-0.4714	-0.4486	0.0141	0.0142	0.0176	0.0204
6	-0.5481	-0.5475	-0.5156	-0.4892	0.0747	0.0746	0.0711	0.0685
7	-0.6608	-0.6602	-0.6274	-0.5998	-0.0033	-0.0033	-0.0063	-0.0080
8	-0.5294	-0.5290	-0.5110	-0.4945	0.2299	0.2293	0.1991	0.1770
9	-0.5408	-0.5404	-0.5217	-0.5046	0.2508	0.2501	0.2204	0.1987
10	-0.5620	-0.5610	-0.5242	-0.5024	0.5725	0.5679	0.4151	0.3444
11	-0.6573	-0.6563	-0.6177	-0.5907	0.5288	0.5242	0.3741	0.3067
12	-0.6471	-0.6461	-0.6049	-0.5767	0.5629	0.5584	0.4131	0.3475
13	-0.7656	-0.7647	-0.7220	-0.6882	0.4898	0.4853	0.3420	0.2804
14	-0.8020	-0.8011	-0.7568	-0.7210	0.4913	0.4869	0.3462	0.2856
15	-0.6003	-0.6022	-0.6386	-0.6324	1.0308	1.0169	0.6319	0.4917
16	-0.4780	-0.4810	-0.5537	-0.5645	1.4367	1.4166	0.8657	0.6664
17	-0.4687	-0.4716	-0.5408	-0.5496	1.4496	1.4296	0.8833	0.6867
18	-0.5443	-0.5474	-0.6178	-0.6221	1.4122	1.3922	0.8481	0.6551
19	-0.5974	-0.6008	-0.6791	-0.6852	1.3758	1.3557	0.8071	0.6134
20	-0.5358	-0.5393	-0.6258	-0.6395	1.6594	1.6359	0.9854	0.7506
21	-0.5426	-0.5461	-0.6342	-0.6487	1.6558	1.6323	0.9813	0.7462
22	-0.6471	-0.6491	-0.6902	-0.6867	1.7038	1.6804	1.0272	0.7871
23	-0.6593	-0.6614	-0.7047	-0.7020	1.6996	1.6761	1.0230	0.7832

Table 8.3: Data and results of the meta-analysis of 53 studies on nicotine replacement therapy for smoking cessation by Stead et al. (2008) data. The subscripts T and C refer to the treatment and control arms of the studies. The columns headed n_T and n_C are the sample sizes, and x_T and x_C are the numbers of events in each study. The columns headed ϕ, v and ϕ_{cum} are the log-relative risks, their variances and cumulative effects. The next five columns (Z_0, B_0, $z_{0.41}$, $-B_{0.41}$, $+B_{0.41}$) are the results of SMA; Z_c are the value of the Z-test for CMA with target value c, $-B_c$ and $+B_c$ the lower and upper Pocock's boundaries for the SMA. The next two columns ($L_{(0,-1.64)}$, $L_{(-0.934,\pm1.96)}$) are the results of penalized Z-test; $L_{(c,b)}$ are the values of penalized Z-test with target value c and b is the critical value. Finally the last eight columns are the results of Gombay test for REM. $GH_{(c,b)}$, $GDL_{(c,b)}$, $GMP_{(c,b)}$, and $GREML_{(c,b)}$ are the results of Gombay test for REM based on $\hat{\tau_H^2}$, $\hat{\tau_{DL}^2}$, $\hat{\tau_{MP}^2}$ and $\hat{\tau_{REML}^2}$, respectively, c is the target value and b i the bootstrap critical values or b1 and b2 are the lower and upper critical values, respectively.

S/N	Author (Year)	xT	nT	xC	nC	$\hat{\phi}$	v	ϕ_{cum}	$z_{0.41}$	$-B_{0.41}$	$+B_{0.41}$	$L_{(0.41,\pm1.96)}$
1	Puska (1979)	29	116	21	113	0.3	0.06	0.30	-0.4343	-3.32	3.32	-0.3895
2	Malcom (1980)	6	73	3	121	1.20	0.48	0.52	0.0195	-3.33	3.33	0.1889
3	Fagerstrom (1982)	30	50	23	50	0.27	0.04	0.33	-0.4977	-3.38	3.38	-0.2360
4	Fee (1982)	23	180	15	172	0.38	0.10	0.34	-0.4978	-3.15	3.15	-0.3017
5	Jarvis (1982)	22	58	9	58	0.89	0.12	0.41	0.0237	-3.24	3.24	0.0786
6	Br Thor Society (1983)	39	410	111	1208	0.03	0.03	0.32	-0.9410	-2.82	2.82	-0.6518
7	Russell (1983)	81	729	78	1377	0.67	0.02	0.42	-0.0843	-2.67	2.67	0.0199
8	Fegerstrom (1984)	28	96	5	49	1.05	0.20	0.46	0.1909	-2.80	2.80	0.2231
9	Hjalmarson (1984)	31	106	16	100	0.60	0.08	0.47	0.3709	-2.83	2.83	0.3056
10	Jamronzik (1984)	10	101	8	99	0.20	0.20	0.45	0.2687	-2.84	2.84	0.2172
11	Killen (1985)	16	44	6	20	0.19	0.16	0.44	0.1205	-2.88	2.88	0.1060
12	Clavel (1985)	24	205	6	222	1.47	0.20	0.50	0.5484	-2.81	2.81	0.4427
13	Hall (1985)	18	41	10	36	0.46	0.10	0.49	0.5521	-2.86	2.86	0.4231
14	Schneider (1985A)	9	30	6	30	0.41	0.21	0.48	0.5311	-2.87	2.87	0.3940
15	Schneider (1985B)	1	13	3	23	-0.53	1.21	0.47	0.4617	-2.88	2.88	0.3375

Table 8.4: Data and results of the meta-analysis of 53 studies on nicotine replacement therapy for smoking cessation by Stead et al. (2008) data cont'd

S/N	Author (Year)	xT	nT	xC	nC	$\hat{\phi}$	v	ϕ_{cum}	$z_{0.41}$	$-B_{0.41}$	$+B_{0.41}$	$L_{(0.41,\pm1.96)}$
16	Page (1986)	9	93	13	182	0.30	0.17	0.46	0.4114	-2.83	2.83	0.2923
17	Campbell (1987)	13	424	9	412	0.34	0.18	0.46	0.3681	-2.73	2.73	0.2522
18	Hall (1987)	30	71	14	68	0.72	0.08	0.47	0.6112	-2.79	2.79	0.3981
19	Roto (1987)	19	54	7	60	1.10	0.16	0.0.50	0.8982	-2.81	2.81	0.5880
20	Areechon (1988)	56	99	37	101	0.43	0.02	0.49	0.8750	-2.80	2.80	0.5591
21	Harackiewicz (1988)	12	99	7	52	-0.10	0.20	0.47	0.6770	-2.81	2.81	0.4336
22	Huber (1988)	13	54	11	60	0.27	0.13	0.46	0.5953	-2.81	2.81	0.3783
23	Llivina (1988)	61	113	28	103	0.69	0.03	0.48	0.9425	-2.79	2.79	0.5818
24	Tonnesen (1988)	23	60	12	53	0.53	0.09	0.49	0.9854	-2.81	2.81	0.6154
25	Blondal (1989)	30	92	22	90	0.29	0.06	0.47	0.8434	-2.79	2.79	0.5290
26	Garcia (1989)	21	68	5	38	0.85	0.21	0.48	0.9498	-2.81	2.81	0.5961
27	Gilbert (1989)	11	112	9	111	0.19	0.18	0.47	0.8673	-2.78	2.78	0.5494
28	Huges (1989)	23	210	6	105	0.65	0.20	0.48	0.9158	-2.76	2.76	0.5864
29	Huges	15	59	5	19	-0.03	0.20	0.47	0.7561	-2.79	2.79	0.4837
30	Killen (1990)	129	600	112	617	0.17	0.01	0.44	0.2013	-2.65	2.65	0.1446
31	Nakamura (1990)	13	30	5	30	0.96	0.21	0.45	0.3321	-2.71	2.71	0.2259
32	Richmond (1990)	17	200	14	150	-0.09	0.12	0.43	0.1108	-2.70	2.70	0.1055
33	Campbell (1991)	21	107	21	105	-0.02	0.08	0.42	-0.1543	-2.72	2.72	-0.0601
34	Jasen (1991)	49	211	19	82	0.00	0.06	0.40	-0.4769	-2.71	2.71	-0.2890
35	Ockene (1991)	40	402	33	420	0.24	0.05	0.40	-0.6113	-2.65	2.65	-0.4078
36	Segnan (1991)	22	294	37	629	0.24	0.07	0.39	-0.7043	-2.64	2.64	-0.4995
37	Clavel-Chapelon (1992)	47	481	42	515	0.18	0.04	0.38	-0.9005	-2.61	2.61	-0.6720

Table 8.5: Data and results of the meta-analysis of 53 studies on nicotine replacement therapy for smoking cessation by Stead et al. (2008) data cont'd

S/N	Author (Year)	xT	nT	xC	nC	$\hat{\phi}$	v	ϕ_{cum}	$z_{0.41}$	$-B_{0.41}$	$+B_{0.41}$	$L_{(0.41,\pm1.96)}$
38	Mc Govern (1992)	51	146	40	127	0.10	0.03	0.37	-1.2230	-2.65	2.65	-0.9053
39	Mori (1992)	30	178	22	186	0.35	0.07	0.37	-1.2443	-2.65	2.65	-0.9506
40	Nebot (1992)	5	106	13	319	0.15	0.26	0.36	-1.2774	-2.65	2.65	-1.0092
41	Pirie (1992)	75	206	50	211	0.43	0.02	0.37	-1.2200	-2.64	2.64	-0.9961
42	Zelman (1992)	23	58	18	58	0.25	0.06	0.36	-1.3056	-2.67	2.67	-1.1010
43	Nieura (1994)	5	84	4	89	0.28	0.43	0.36	-1.3196	-2.66	2.66	-1.1390
44	Fortmann (1995)	110	552	84	522	0.21	0.02	0.35	-1.5627	-2.58	2.58	-1.3696
45	Gross (1995)	37	131	6	46	0.77	0.16	0.36	-1.4740	-2.63	2.63	-1.2979
46	Herrera (1995)	30	76	13	78	0.86	0.08	0.37	-1.2630	-2.64	2.64	-1.0753
47	Hall (1996)	13	424	9	412	0.34	0.18	0.36	-1.5600	-2.64	2.64	-1.2521
48	Niaura (1999)	1	31	2	31	-0.69	1.44	0.36	-1.5970	-1.65	1.65	-1.2826
49	Villa (1999)	11	21	10	26	0.31	0.10	0.35	-1.6252	-1.66	1.66	-1.3275
50	Garvey (2000)	75	405	17	203	0.79	0.06	0.37	-1.4095	-1.61	1.61	-1.1109
51	Cooper (2005)	17	146	15	147	0.13	0.11	0.36	-1.4955	-1.62	1.62	-1.2008
52	Ahluwailia (2006)	53	378	42	377	0.23	0.04	0.36	-1.6188	-1.58	1.58	-1.3217
53	Moolchan (2008)	8	46	2	40	1.25	0.58	0.36	-1.5596	-1.62	1.62	-1.2734

Table 8.6: Data and results of the meta-analysis of 53 studies on nicotine replacement therapy for smoking cessation by Stead et al. (2008) data data cont'd

S/N	$GH_{(0;0.45)}$	$GDL_{(0;0.45)}$	$GMP_{(0;0.45)}$	$GREML_{(0;0.44)}$	$GH_{(0.41;-0.32,0.46)}$	$GDL_{(0.41;-0.33,0.45)}$	$GMP_{(0.41;-0.31,0.45)}$	$GREML_{(0.41;-0}$
1	0.1446	0.1446	0.1456	0.1418	-0.0597	-0.0597	-0.0601	-0.0585
2	0.2242	0.2242	0.2249	0.2223	0.0027	0.0027	0.0021	0.0044
3	0.2695	0.2695	0.2711	0.2650	-0.0684	-0.0684	-0.0696	-0.0647
4	0.3102	0.3102	0.3119	0.3054	-0.0684	-0.0684	-0.0696	-0.0650
5	0.4132	0.4132	0.4150	0.4079	0.0033	0.0033	0.0022	0.0063
6	0.3607	0.3607	0.3620	0.3570	-0.1293	-0.1293	-0.1321	-0.1213
7	0.5615	0.5615	0.5668	0.5468	-0.0116	-0.0116	-0.0120	-0.0102
8	0.6130	0.6130	0.6181	0.5989	0.0262	0.0262	0.0256	0.0281
9	0.6683	0.6683	0.6736	0.6537	0.0509	0.0509	0.0504	0.0525
10	0.6666	0.6666	0.6719	0.6520	0.0369	0.0369	0.0365	0.0382
11	0.6622	0.6622	0.6675	0.6475	0.0166	0.0166	0.0162	0.0177
12	0.7333	0.7333	0.7383	0.7191	0.0753	0.0753	0.0748	0.0770
13	0.7552	0.7552	0.7603	0.7409	0.0758	0.0758	0.0753	0.0774
14	0.7635	0.7635	0.7686	0.7492	0.0729	0.0729	0.0725	0.0744
15	0.7567	0.7567	0.7618	0.7422	0.0634	0.0634	0.0630	0.0647
16	0.7626	0.7626	0.7678	0.7482	0.0565	0.0565	0.0561	0.0577
17	0.7685	0.7685	0.7737	0.7541	0.0506	0.0506	0.0502	0.0516
18	0.8271	0.8271	0.8326	0.8119	0.0840	0.0840	0.0838	0.0845
19	0.8795	0.8795	0.8849	0.8644	0.1234	0.1234	0.1231	0.1240
20	0.9304	0.9304	0.9369	0.9124	0.1202	0.1202	0.1200	0.1208
21	0.9131	0.9131	0.9196	0.8948	0.0930	0.0930	0.1200	0.0933
22	0.9156	0.9156	0.9222	0.8974	0.0818	0.0818	0.0929	0.0821
23	1.0038	1.0038	1.0116	0.9820	1.1295	0.1295	0.0817	0.1278
24	1.0270	1.0276	1.0354	1.0056	1.1354	0.1354	0.0817	0.1336

Table 8.7: Data and results of the meta-analysis of 53 studies on nicotine replacement therapy for smoking cessation by Stead et al. (2008) data cont'd

S/N	$GH_{(0;0.45)}$	$GDL_{(0;0.45)}$	$GMP_{(0;0.45)}$	$GREML_{(0;0.44)}$	$GH_{(0.41;-0.31,0.45)}$	$GDL_{(0.41;-0.33,0.45)}$	$GMP_{(0.41;-0.31,0.45)}$	$GREML_{(0.41;-0}$
25	1.0337	1.0337	1.0416	1.0115	1.1158	0.1158	0.1301	0.1145
26	1.0569	1.0569	1.0647	1.0349	1.1305	0.1305	0.1360	0.1292
27	1.0547	1.0547	1.0625	1.0326	1.1191	0.1191	0.1164	0.1178
28	1.0700	1.0700	1.0778	1.0480	1.1258	0.1258	0.1309	0.1244
29	1.0568	1.0568	1.0646	1.0347	1.1039	0.1039	0.1196	0.1024
30	1.0418	1.0418	1.0493	1.0208	1.0277	0.0277	0.1263	0.0321
31	1.0674	1.0674	1.0748	1.0467	0.0456	0.0456	0.0438	0.0503
32	1.0491	1.0491	1.0565	1.0283	0.0152	0.0152	0.0135	0.0197
33	1.0300	1.0302	1.0375	1.0094	-0.0212	-0.0212	-0.0230	-0.0164
34	1.0079	1.0079	1.0152	0.9875	-0.0655	-0.0655	-0.0676	-0.0600
35	1.0128	1.0128	1.0201	0.9921	-0.0840	-0.0840	-0.0861	-0.0781
36	1.0181	1.0181	1.0255	0.9974	-0.0967	-0.0967	=0.0989	-0.0908
37	1.0174	1.0174	1.0248	0.9966	-0.1237	-0.1237	-0.1261	-0.1170
38	1.0052	1.0052	1.0124	0.9848	-0.1682	-0.1682	-0.1713	-0.1598
39	1.0194	1.0194	1.0268	0.9988	-0.1709	-0.1709	-0.1740	-0.1625
40	1.0201	1.0201	1.0274	0.9994	-0.1755	-0.1755	-0.1785	-0.1671
41	1.0638	1.0638	1.0720	1.0407	-0.1676	-0.1676	-0.1704	-0.1598
42	1.0690	1.0690	1.0772	1.0458	-0.1893	-0.1793	-0.1822	-0.1714
43	1.0703	1.0703	1.0785	1.0471	-0.1813	-0.1813	-0.1841	-0.1734
44	1.0782	1.0782	1.0867	1.0544	-0.2147	-0.2147	-0.2183	-0.2047
45	1.0978	1.0978	1.1063	1.0741	-0.2025	-0.2025	-0.2061	-0.1924
46	1.1397	1.1397	1.1482	1.1157	-0.1735	-0.1735	-0.1771	-0.1636
47	1.1163	1.1163	1.1247	1.0926	-0.2143	-0.2143	-0.2181	=0.2038

Table 8.8: Data and results of the meta-analysis of 53 studies on nicotine replacement therapy for smoking cessation by Stead et al. (2008) data cont'd

S/N	$GH_{(0;0.45)}$	$GDL_{(0;0.45)}$	$GMP_{(0;0.45)}$	$GREML_{(0;0.44)}$	$GH_{(0.41;-0.31,0.45)}$	$GDL_{(0.41;-0.33,0.45)}$	$GMP_{(0.41;-0.31,0.45)}$	$GREML_{(0.41;-0}$
48	1.1124	1.1124	1.1208	1.0886	-0.2194	-0.2194	-0.2231	-0.2090
49	1.1191	1.1191	1.1276	1.0953	-0.2233	-0.2233	-0.2271	-0.2130
50	1.1643	1.1643	1.1730	1.1398	-0.1936	-0.1936	-0.1972	-0.1838
51	1.1621	1.1621	1.1708	1.1376	-0.2054	-0.2054	-0.2090	-0.1956
52	1.1682	1.1682	1.1770	1.1434	-0.2225	-0.2224	-0.2261	-0.2121
53	1.1786	1.1786	1.1874	1.1541	-0.2142	-0.2142	-0.2180	-0.2038