RANDOMISED CLINICAL TRIALS: STATISTICAL INVESTIGATIONS OF BIAS, INEFFICIENCY AND MISINTERPRETATION

Ishani Manjula Schou

Department of Statistics

Faculty of Science and Engineering

Macquarie University

2016

This thesis is submitted in total fulfilment of

the requirements for the degree of

Doctor of Philosophy.

Abstract

Randomised clinical trials (RCTs) compare treatment interventions using the health outcomes of individuals assigned to their treatment at random. RCTs are the gold standard for comparing treatment efficacy, but many factors in their design, conduct and analysis can lead to bias, inefficiency or misinterpretation. This thesis by publication presents statistical investigations of three such areas.

The first area relates to potential bias from early stopping of RCTs. Some researchers have claimed that early stopping of RCTs based on interim analyses leads to overestimation of the treatment effect and that this is particularly problematic for meta-analyses that synthesise the results of multiple studies. This thesis presents extensive theoretical and simulation studies of this potential source of bias. It is concluded that early stopping is not a substantive source of bias for meta-analyses of RCTs.

The second area relates to the potential for misinterpretation of RCT subgroup analyses, particularly subgroups defined by geographical region in global studies. Subgroup analysis principles require a significant test of interaction to conclude heterogeneity of subgroup treatment effects. However, overly optimistic expectations of treatment effect homogeneity often lead to over-interpretation of apparent differences between subgroups. This thesis proposes a suite of graphical analyses that supplement a test of interaction with a visual assessment of the extent to which chance can explain the observed differences between subgroups. An open-source software package for the R computing environment is presented.

The third area relates to efficient design of RCTs having several treatments compared to a common control. Standard balanced designs have equal numbers of individuals on each treatment, but are inefficient in this context. This thesis considers efficient unbalanced designs that minimise variance or maximise power. New results in optimal design theory, and some guidelines for the efficient planning of RCTs having several treatments, are presented.

Statement of Candidate

I certify that, except where acknowledged, the material presented in this thesis is original, and has not been submitted in whole or in part for a degree at any university or institution.

I also certify that this thesis has been written by me, and in totality, my contribution was at least 90% of the total effort required to conduct and complete this research. Specifically, the research conducted as part of this degree contributed towards four papers, three of which were co-authored with my supervisor, Professor Ian C. Marschner. In each of these cases, the contribution of the co-author was to assist with conception of the methods, and provide general supervision and feedback on the research and writing. Additionally, a fourth paper and a software package of which I was the sole author, are also included in this thesis.

Finally, I certify that all information sources and literature used are referenced in the thesis.

Signature:

Name: Ishani Manjula Schou

Date:

Acknowledgements

To my supervisor and guru, Professor Ian Marschner, my deepest gratitude for his generosity of time, his patient instruction, his consistent support, his droll observations and his insightful comments, with all of which he guided me on this journey of learning that has brought me great joy.

To my associate supervisor, Professor John Simes (NHMRC Clinical Trials Centre, University of Sydney), my grateful thanks for his thoughtful considerations of the applicability of my research from the perspective of evidence based medicine and clinical decision making.

Til min kjæreste Erik, hjertelig takk for hans urokkelige støtte, hans uendelige tålmodighet og for hans standhaftige tro på mine evner til å gjøre dette ferdig.

Publications

Peer-reviewed articles

Schou, I. M. and I. C. Marschner (2013). Meta-analysis of clinical trials with early stopping: an investigation of potential bias. *Statistics in Medicine* **32**: 4859–4874.

Schou, I. M. and I. C. Marschner (2015). Methods for exploring treatment effect heterogeneity in subgroup analysis: an application to global clinical trials. *Pharmaceutical Statistics* **14**: 44–55.

Submitted article

Schou, I. M. and I. C. Marschner (2016). Design of clinical trials involving multiple hypothesis tests with a common control. Submitted to *Biometrical Journal*.

Manuscript ready for submission

Schou, I. M. (2016). subgroup: A package for exploring treatment effect heterogeneity in subgroup analysis.

Software

Schou, I.M. subgroup: Methods for exploring treatment effect heterogeneity in subgroup analysis of clinical trials. R package version 1.1 2014. Available at: http://CRAN.R-project.org/package=subgroup.

Abstracts

Schou, M. and I. Marschner. Biases in clinical trials with sequential monitoring. *MRC-HTMR Clinical Trials Methodology Conference*. Bristol, UK, October, 2011. Published in: *Trials* 2012(**12**), Suppl 1: A50.

Marschner, I. and M. Schou. Evidence synthesis when clinical trials report early stopping. *2012 Scientific Symposium of the EQUATOR Network*. Freiburg, Germany, 2012.

Schou, I. M. and I. C. Marschner. Methods for exploring treatment effect heterogeneity in subgroup analysis: an application to global clinical trials. *International Society for Clinical Biostatistics 34th Annual Conference*. Munich, Germany, 2013.

Schou, I. M. and I. C. Marschner. Optimal design of clinical trials involving multiple tests with a common control. *International Society for Clinical Biostatistics 36th Annual Conference*. Utrecht, Netherlands, 2015.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis is concerned with statistical methodology for the design, monitoring and analysis of randomised clinical trials (RCTs). Although RCTs are ubiquitous in the medical literature and are the gold standard for comparing treatment efficacy, many complexities can arise in practice that may lead to bias, inefficiency or misinterpretation. Resource limitations and a need to ensure that no more patients than necessary are exposed to experimental and unproven therapies, are two important considerations necessitating clinical trials to be designed optimally. In addition, an ethical imperative exists that new treatments with proven efficacy are made available to patients in a timely manner. This induces a need for trials requiring long-term follow-up, in particular those trials that can have significant clinical impact such as mortality studies, to be monitored at interim time-points and potentially stopped early. This adds complexity to the design and analysis of a clinical trial and requires specialised methods of statistical inference. Given these challenges, maximising the use of available data and evidence, either through evaluations of the efficacy of new therapies using subgroup analyses or through the synthesis of evidence from multiple trials addressing the same question is frequently encountered in medical literature.

The research presented addresses such issues with reference to three specific topics relevant to clinical trials research: investigating the implications of including in meta-analyses the results from trials subjected to interim monitoring and early stopping; understanding the play of chance when interpreting differences in subgroup-specific treatment effects with a particular focus on subgroups defined by geographical regions in multi-country trials;

and optimising efficiency through unbalanced designs of single-control multiple-comparator clinical trials. In the remainder of this chapter, an overview of the thesis is provided along with a summary of the motivation and main lines of research on the above topics.

1.1 Overview

This thesis is presented as a "thesis by publication" and consists of six chapters. A brief outline of each of the three topics discussed above is presented in this introductory chapter along with some review of existing methodology. This chapter also describes the research motivation behind each of these chapters and the contribution it makes to the existing body of knowledge. The four chapters that constitute the main research were each written as journal articles and are reproduced in this thesis in a common format.

Chapter 2 and its associated Appendix 2.A present the investigations on bias in meta-analyses involving trials subjected to interim monitoring and has been published in the journal *Statistics and Medicine* as Schou and Marschner (2013). The content of Chapter 3 investigates potential misinterpretation of subgroup analyses, particularly those from multi-country studies, and has been published in the journal *Pharmaceutical Statistics* as Schou and Marschner (2015). Chapter 4 contains details of a software package that implements the methodology presented in Chapter 3, and is presented as a manuscript ready for submission for peer review. The documentation associated with this open-source package for the R computing environment is presented in the Appendix to the chapter. Chapter 5, which has been submitted for peer review, presents the methodological findings relating to unbalanced designs of single-control multiple-comparator clinical trials in which multiple hypothesis tests are conducted involving a common control. Finally, Chapter 6 is a concluding chapter that consolidates the content of this thesis and presents areas for future research.

The remaining sections of this chapter review the background and motivation for the primary areas of study in this thesis.

1.2 Meta-analyses and interim monitoring

Design and analysis of group sequential clinical trials are generally well understood and implemented in clinical research (Whitehead, 1997; Jennison and Turnbull, 2000). This methodology allows the accrued data of an ongoing trial to be analysed sequentially while controlling the type I error rate at the desired overall significance level which otherwise would be inflated as a result of the multiplicity of testing. Typically, a study is designed with a stopping boundary in mind, such as the O'Brien-Fleming or Haybittle-Peto boundaries, and the analysis conducted based on the critical values defined by these boundaries. Having the opportunity to analyse the data during the course of a study has the benefit of making effective new therapies available faster, or stopping a trial that is unlikely to result in a worthwhile benefit or compromises patient safety.

It has long been acknowledged that trials stopping early for benefit may have a naive estimate of treatment effect that is inflated. A remedy for this was first suggested by Whitehead (1986), using a bias-adjusted maximum likelihood estimate, and this approach was further developed from a computational and inferential perspective by Todd et al. (1996). Subsequently, many other methods of bias adjustment were also suggested as discussed further in Chapter 6. The potential for such bias has naturally led researchers to question whether trials that stop early for benefit will introduce bias into meta-analyses. It is clear from the literature that some researchers strongly advocate that results from trials that truncate early, and consequently results of meta-analyses that include truncated studies, lead to overestimates of treatment effect (Montori et al., 2005; Bassler et al., 2008; Bassler et al., 2010). Some authors have argued that systematic reviews should explore truncation as an explanation for heterogeneity in meta-analyses (Bassler et al., 2007), which is bound to lead systematic reviewers to consider excluding truncated studies. Indeed, some researchers have advocated that sensitivity analyses be conducted excluding truncated studies (Bassler et al., 2013). Other researchers however, oppose this thinking on an intuitive level backed up by statistical reasoning (Goodman, 2008; Berry, Carlin, and Connor, 2010), and through the implementation of simulation studies (Green, Fleming, and Emerson, 1987; Todd, 1997).

The conflicting standpoints of researchers on this question provided the impetus for the research presented in Chapter 2. The focus of this chapter was to investigate the consequences of including only non-truncated studies in evidence synthesis. The chapter begins with a theoretical quantification of the estimation and information biases in the special case of a trial subjected to a single interim analysis, supplemented by simulation studies of trials subjected to more than one interim analysis. This leads to the conclusion that excluding truncated studies from meta-analyses leads to underestimation of the treatment effect and overestimation of the information associated with the treatment effect. This chapter concludes that early stopping is not a substantive source of bias for meta-analyses and that both truncated and non-truncated trials should be included in evidence synthesis.

Subsequent work on the theoretical quantification of biases in trials subjected to interim monitoring is presented in the concluding chapter. This provides some direction for future research in the area of bias resulting from early stopping.

1.3 Subgroup analyses

Subgroup analysis principles have long been an important area of statistical methodology for RCTs, particularly with a focus on the potential for multiple testing problems associated with the conduct of many subgroup analyses. The related issue of interpreting observed differences in treatment effects across subgroups, has recently been widely debated in the literature with a particular focus on subgroup analyses involving country-specific analyses in multi-country studies.

With increasing globalisation of drug development programs which enable faster entry of new therapies to market and allows patients across the globe to have access to new treatments, multi-country randomised clinical trials are increasingly common. However, it is seldom that a test of interaction is adequately powered to assess treatment effect heterogeneity across countries or geographic regions. As a result, interpreting apparent treatment effect differences can be difficult, and may lead to undue speculation about the causes of

such differences. In particular, recent debate has centred on potential treatment effect heterogeneity across country or region of randomisation, and the interpretation of results when some countries seemingly favour the control treatment (Wallentin et al., 2009; Wedel et al., 2001). In some instances, this has led to further analysis of the trial data by regulatory authorities (FDA, 2010) and exploration of the study conduct to identify any underlying factors that may have contributed to these differences (Serebruany, 2010). On the other hand, some authors have emphasised the contribution that chance variation can play (Buyse and Marschner, 2011). In this environment, a joint workshop titled "*Ensuring Quality and Balancing Risks for Multiregional Clinical Trials: Clinical, Regulatory, and Ethical Factors*" was organised by the Drug Information Association (DIA), the Food and Drug Administration (FDA), and the pharmaceutical industry in Washington DC. The workshop had the intention of discussing, understanding and coming to a common consensus on some issues pertaining to multi-country randomised clinical trials. A central area of discussion was the issue of anticipating and understanding the magnitude of treatment effect differences that can arise merely as an artefact of chance (Ibia and Binkowitz, 2011).

Chapter 3 is therefore motivated by further research into the problem of quantifying the nature and magnitude of chance variation in subgroup-specific treatment effects. It provides a set of graphical analysis tools by which a researcher can compare the expected variation with the observed variation. Although the motivation for this investigation arose from subgroup analyses of multi-country studies, where the subgroups were defined by the country or region of randomisation, the proposed methodology is equally applicable to all subgroup analyses. Indeed, as the financial burden and regulatory requirements of conducting clinical trials are substantial, researchers often include pre-specified subgroup analyses to glean as much information as possible about the effect of a new therapy in different types of patients. The methodology proposed in Chapter 3 is considered to be supplemental to a test of interaction, and describes graphical non-inferential methods by which to compare observed treatment differences with those that could have occurred due to chance, that is, under an assumption of treatment effect homogeneity across subgroups.

Implementation of the methodology proposed in Chapter 3 is computationally intensive. Chapter 4 discusses in further detail the implementation of these methods through the R software package **subgroup** which has been included in the Comprehensive R Archive Network (CRAN). It describes how this package can be utilised by researchers at the design phase of a clinical trial to understand the range of treatment effects that might be anticipated as a result of chance, and thereafter during the analysis phase to compare the observed differences with the expected differences under an assumption of homogeneity of treatment effects across subgroups. The package allows further flexibility by allowing simulation based calculations to be implemented when the number of subgroups is large. The default graphics provide a useful non-inferential tool for understanding chance variation in treatment effects across subgroups.

1.4 Optimal design

Randomised clinical trials in which a single control is compared with multiple comparators are often encountered in the medical literature. This includes situations where the control treatment is a placebo, as in trials in which a placebo is to be compared with multiple experimental treatments, or an active treatment, as in trials where a combination therapy is to be compared with each of its constituent mono-therapies. Unification of these multiple comparisons under a single protocol offers efficiency in trial conduct. That is, consolidation of trial related activities such as ethics approvals, investigator site initiations, database development and other activities may offer efficiencies leading to faster or cheaper study conduct. It is evident from the literature that a balanced design, that is, one in which an equal number of patients is randomised to each group is the preferred design in most single-control multiple-comparator clinical trials. However, a more efficient trial design can be achieved with an optimal design of these trials through the unbalanced allocation of patients. In classical experimental design, this optimisation may be achieved through the minimisation of some variance measure (Atkinson and Donev, 1992; Hedayat, Jacroux, and Majumdar, 1988). In the clinical trials context, it has been suggested that the optimisation involve maximising

the power of the study (Marschner, 2007). Weighted versions of variance optimal measures have also been proposed (Morgan and Wang, 2010), including in the context of clinical trials (Zhu and Wong, 2000; Wong and Zhu, 2008). Nevertheless, the design efficiency that can be gained through unbalanced allocation of patients in single-control multiple-comparator trials is often overlooked. Chapter 5 will explore such unbalanced designs. The first part of Chapter 5 will explore three variance optimal designs based on the D-, A- and E-optimality criteria, with or without weighting. These will be unified under a single form for a general model that allows heteroscedasticity and continuous or binary outcomes. This unification will allow the sensitivity of the design to the chosen variance optimality criterion to be evaluated through a comparison of the way in which each method allocates the available resources to the control and comparator arms. As these variance optimality methods focus on estimation precision, the second part of Chapter 5 considers optimisation of power as the more suitable approach for clinical trials where testing is usually the focus. However, as unbalanced designs which optimise power are complex, an investigation into whether any of these variance optimal designs can be used as satisfactory approximations to the power optimal designs will be conducted. The ultimate goal of this research is to provide some guidelines on how resources may be allocated in an unbalanced fashion in single-control multiple-comparator trials. This will equip researchers with some simple rules of thumb for determining an unbalanced design when multiple hypothesis tests are to be conducted using a common control.

Chapter 2

Meta-analysis of clinical trials with early stopping

Published as:

Schou, I. M. and I. C. Marschner (2013). Meta-analysis of clinical trials with early stopping: an investigation of potential bias. *Statistics in Medicine* **32**: 4859–4874.

Abstract

Clinical trials that stop early for benefit have a treatment difference that overestimates the true effect. The consequences of this fact have been extensively debated in the literature. Some researchers argue that early stopping, or truncation, is an important source of bias in treatment effect estimates, particularly when truncated studies are incorporated into meta-analyses. Such claims are bound to lead some systematic reviewers to consider excluding truncated studies from evidence synthesis. We therefore investigated the implications of this strategy, by examining the properties of sequentially monitored studies conditional on reaching the final analysis. As well as estimation bias, we studied information bias measured by the difference between standard measures of statistical information, such as sample size, and the actual information based on the conditional sampling distribution. We found that excluding truncated studies leads to underestimation of treatment effects and overestimation of information. Importantly, the information bias increases with the estimation bias, meaning that greater estimation bias is accompanied by greater overweighting in a meta-analysis. Simulations of meta-analyses confirmed that the bias from excluding truncated studies can be substantial. In contrast, when meta-analyses included truncated studies, treatment effect

estimates were essentially unbiased. Previous analyses comparing treatment effects in truncated and non-truncated studies are shown not to be indicative of bias in truncated studies. We conclude that early stopping of clinical trials is not a substantive source of bias in meta-analyses and recommend that all studies, both truncated and non-truncated, be included in evidence synthesis.

Keywords: Bias; clinical trial; interim analysis; meta-analysis; truncation

2.1 Introduction

Group sequential designs are widely used for interim monitoring of randomised clinical trials (RCTs) and often allow such studies to be stopped early, or truncated, due to an apparent treatment benefit. Since interim analyses that test for a treatment benefit will elevate the probability of a type I error, various methods have been developed to control the effects of multiple testing. These have included commonly used sequential stopping rules such as the O'Brien-Fleming and Haybittle-Peto boundaries, as well as the Lan-DeMets error spending generalisations and other approaches (Jennison and Turnbull, 2000; Whitehead, 1997). Such methods are generally well understood and appropriately implemented in practice.

Even when an appropriate statistical stopping rule has been used to stop a study early, there can still be bias in the treatment effect estimate and its associated standard error, confidence interval and p-value. This is because random fluctuations favouring the experimental treatment may lead to truncation of an RCT, which in turn leads to overestimation of the treatment effect. The potential for such bias has been known for decades (Hughes and Pocock, 1988; Whitehead, 1986; Pocock and Hughes, 1989) and various methods of analysis following termination of a sequential trial have been proposed (Jennison and Turnbull, 2000; Whitehead, 1997). Nonetheless, such methods are generally less well understood than methods for controlling type I error and are rarely used in practice when reporting the results of RCTs.

In recent years there has been a renewed focus on early stopping of RCTs. Some researchers

have made strong claims that this is a source of bias in the literature and that the results of truncated RCTs should be viewed with scepticism (Montori et al., 2005; Bassler et al., 2008; Bassler et al., 2010). The potential effect of this bias on systematic reviews has been of particular concern (Bassler et al., 2007). Such arguments have been countered by other researchers, who have claimed that truncation of RCTs does not lead to substantive bias (Goodman, 2008; Berry, Carlin, and Connor, 2010; Freidlin and Korn, 2009). This debate is potentially confusing for systematic reviewers who have to synthesise evidence that often contains truncated RCTs. Although the standard approach has been to incorporate truncated RCTs without any special consideration, claims that early stopping is an important source of bias will inevitably lead some systematic reviewers to consider excluding truncated studies from meta-analyses. We have therefore undertaken an investigation of whether such a strategy is advisable, and how it compares to the standard approach of including truncated studies in meta-analyses.

The effects of early stopping on meta-analyses were studied by Hughes et al. (1992), however, the primary focus of that paper is different to ours. Hughes et al. (1992) focused on the effects of interim monitoring on heterogeneity tests with normal endpoints. They also looked at estimation bias in the context of meta-analysing all studies where all studies have sequential monitoring. Here we look at a range of additional issues. Firstly, we study a mix of fixed and sequential studies and investigate how the relativity of this mix affects the results. We also study a range of alternative meta-analysis strategies, motivated by the prior claims of bias in truncated studies. Thus, we focus on the performance of strategies that involve excluding truncated studies or excluding all sequentially monitored studies. We also compare truncated and non-truncated studies to investigate their expected differences, since such comparisons have been used to infer bias in systematic reviews of truncated studies (Bassler et al., 2010). Finally, we consider extension of the results to binary endpoint contexts.

Our primary approach to investigating the exclusion of truncated RCTs is to consider the statistical properties of sequentially monitored studies conditional on reaching the planned

final analysis. Our investigations indicate that a strategy of excluding truncated RCTs from meta-analyses can lead to substantial bias. This bias arises in both the estimated treatment effect, which we call estimation bias, as well as in the information weights used to aggregate study-specific estimates in meta-analyses, which we call information bias. In contrast, inclusion of truncated RCTs in aggregated treatment effect estimates is found to be virtually unbiased. We therefore conclude that early stopping of clinical trials due to treatment benefit is not an important source of bias in systematic reviews and that inclusion of all studies, both truncated and non-truncated, should remain the standard approach to evidence synthesis.

2.2 Assumptions and notation

2.2.1 Meta-analysis

Consider the comparison of two treatment groups in a single RCT having a fixed total sample size of n individuals, with a proportion γ_j allocated to treatment group j. Let X_{ij} be the outcome variable for individual i in group j. It is assumed that $\{X_{ij}\}$ are independent random variables with means $E(X_{ij}) = \mu_j$, where $\mu_2 = \mu_1 + \delta$, and variances $\text{var}(X_{ij}) = \sigma_j^2$. The parameter δ measures the treatment effect, with $\delta = 0$ corresponding to no difference between the treatments. The estimator of δ is $D = \overline{X}_2 - \overline{X}_1$, the difference in sample means between the two treatment groups, which is assumed to be normally distributed with $E(D) = \delta$ and variance specified by $n\text{var}(D) = \nu^2 = \sigma_1^2/\gamma_1 + \sigma_2^2/\gamma_2$. Thus,

$$D \sim N\left(\delta, I^{-1}\right) \qquad \text{where} \qquad I = n/\nu^2. \tag{2.1}$$

The above model, which would be typical for studies with continuous endpoints, is sufficient for addressing the main issues discussed in this paper. Some discussion of modifications that allow for other types of endpoints, particularly binary endpoints with relative risk or risk difference effect measures, will be provided later in the paper. For now it suffices to say that such modifications do not alter our main conclusions.

Now consider a meta-analysis of M such RCTs indexed by $m = 1,\ldots,M$, with study m having treatment difference estimator D_m and sample size n_m. Then for study weights $\{w_m\}$ with $\sum_m w_m = 1$, the aggregated estimator is

$$\hat{\Delta} = \sum_{m=1}^{M} w_m D_m. \tag{2.2}$$

The weights w_m reflect the amount of statistical information provided by study m, or equivalently the precision of the study-specific estimators D_m. For a fixed effects meta-analysis the analogue of model (2.1) for study m is

$$D_m \sim N\left(\delta, I_m^{-1}\right) \qquad \text{where} \qquad I_m = n_m/\nu_m^2 \tag{2.3}$$

and $\hat{\Delta}$ in (2.2) would be calculated using the information weights

$$w_m = \frac{I_m}{\sum_{h=1}^{M} I_h}. \tag{2.4}$$

Fixed effects meta-analysis will be our primary concern in this paper, however, it is important to consider the robustness of our conclusions for random effects meta-analyses. This is particularly true given that early stopping of RCTs can impact on tests of heterogeneity which are sometimes used to choose between fixed and random effects approaches (Goodman, 2008; Hughes, Freedman, and Pocock, 1992). For a random effects meta-analysis, model (2.3) is used with the generalisation

$$I_m = \left(\frac{\nu_m^2}{n_m} + \tau^2\right)^{-1} \tag{2.5}$$

where $\tau^2 \geq 0$ is the variance of the study-specific treatment effects. With this generalisation the form of the aggregated estimate and the information weights remain as in the fixed effects case specified in (2.2) and (2.4).

Note that since the information weights (2.4) depend on variance parameters ν_m, in practice

I_m would be replaced by an estimate $\hat{I}_m$ using a variance estimate $\hat{v}_m$. In the context of random effects meta-analysis, estimation of τ^2 is also required. In our theoretical investigations we will assume known variance so that the information weights are known and estimation is not necessary, but in our simulation studies of meta-analyses we will allow for estimation of the information weights.

2.2.2 Interim analysis

The previous section assumed that study m has a fixed sample size n_m. Now suppose that study m has a fixed planned sample size n_m^*, but that the actual observed sample size n_m could be smaller due to the conduct of one or more interim analyses. The sample size at interim analysis k for study m is denoted $n_m^{(k)} \leq n_m^*$ and the difference in sample means between the two treatment groups is denoted $D_m^{(k)}$. At each interim analysis a superiority test is conducted of the null hypothesis $H_0 : \delta = 0$ versus the one-sided alternative $H_0 : \delta > 0$, and if H_0 is rejected then the study stops with a conclusion that treatment 2 is more efficacious than treatment 1. Otherwise, the study proceeds to the next interim analysis where the same testing procedure is repeated with a larger sample size, and if the planned final sample size n_m^* is reached then the study stops with a final treatment difference D_m^*. Under this scheme n_m is a random variable with maximum value n_m^*, and its distribution depends on the way in which the interim analyses are carried out.

We assume that the interim analyses are carried out such that H_0 is rejected at analysis k if

$$D_m^{(k)} > b_m^{(k)} {I_m^{(k)}}^{-0.5} \qquad \text{where} \qquad I_m^{(k)} = n_m^{(k)} / v_m^2. \tag{2.6}$$

The constants $b_m^{(k)}$ are the stopping boundaries on the standardised test statistic scale, and are chosen so as to achieve the desired study-wise type I error probability α_m. In this paper we will determine $b_m^{(k)}$ using the standard Lan-DeMets error spending approximation to the O'Brien-Fleming, Haybittle-Peto or Pocock stopping boundaries (Jennison and Turnbull, 2000). This requires the standard assumption that the timing of each interim analysis is not

data dependent, or equivalently, that the sample size $n_m^{(k)}$ is independent of the values of the treatment differences $\{D_m^{(l)}; l \leq k\}$. As mentioned in Section 2.2.1, v_m^2 has to be estimated in practice. This estimation will be undertaken in our simulation studies of meta-analyses, whereas our theoretical investigations will use the simplifying assumption of known variance.

On termination of the study, the final treatment difference D_m will be either D_m^* or one of the interim differences $D_m^{(k)}$, and the final sample size n_m will be either n_m^* or one of the interim sample sizes $n_m^{(k)}$. Meta-analysis then involves aggregating the final study-specific treatment differences D_m using $\hat{\Delta}$ in (2.2), with I_m based on n_m through either the fixed or random effects specifications in (2.3) and (2.5). In principle this aggregation could be undertaken using treatment differences that are adjusted for the sequential analyses (Jennison and Turnbull, 2000; Whitehead, 1997), although that will not be the focus here. Some brief comments on the use of adjusted differences will be made in Section 2.6.

While it will be of interest to study the behaviour of meta-analyses involving both truncated and non-truncated studies, our main aim is to study the effects of excluding truncated studies from meta-analyses. This exclusion amounts to conditioning on non-truncation, that is, conditioning on non-rejection of H_0 at each of the interim analyses. Subject to this conditioning, we will study the statistical properties of D_m and $\hat{\Delta}$ to quantify the extent to which exclusion of truncated studies biases estimation of treatment effects.

2.3 Conditioning on non-truncation

The distribution of D_m conditional on non-truncation can provide insight into the effect of excluding truncated studies from meta-analyses. This conditional distribution can be studied using the joint distribution of the treatment differences at the interim and final analyses. In this section we consider this for the fixed effects meta-analysis model with known variance parameter v_m. In later sections we will also study the meta-analysis estimator $\hat{\Delta}$ using simulation for both the fixed and random effects models with estimated variance.

Consider a study m with up to K_m interim analyses conducted according to the rejection rule (2.6) and a final $(K_m+1)^{\text{th}}$ analysis with treatment difference D_m^*, sample size n_m^* and information $I_m^* = n_m^*/v_m^2$. Let $\mathbf{D}_m = \left(D_m^{(1)}, \ldots, D_m^{(K_m)}, D_m^*\right)$ be the vector of treatment differences at the interim and final analyses, and let $\mathbf{d}_m$ be a vector of length K_m+1 with all elements equal to δ. Then, using the canonical joint distribution of the test statistics in a sequentially monitored study (Jennison and Turnbull, 2000), the distribution of $\mathbf{D}_m$ is a (K_m+1)-dimensional multivariate normal distribution

$$\mathbf{D}_m \sim N\left(\mathbf{d}_m, V_m\right) \qquad \text{where} \qquad V_m = \begin{bmatrix} {I_m^{(1)}}^{-1} & {I_m^{(2)}}^{-1} & \ldots & {I_m^{(K_m)}}^{-1} & {I_m^*}^{-1} \\ {I_m^{(2)}}^{-1} & {I_m^{(2)}}^{-1} & \ldots & {I_m^{(K_m)}}^{-1} & {I_m^*}^{-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ {I_m^{(K_m)}}^{-1} & {I_m^{(K_m)}}^{-1} & \ldots & {I_m^{(K_m)}}^{-1} & {I_m^*}^{-1} \\ {I_m^*}^{-1} & {I_m^*}^{-1} & \ldots & {I_m^*}^{-1} & {I_m^*}^{-1} \end{bmatrix}. \quad (2.7)$$

Conditional on non-truncation, that is conditional on the event

$$\mathscr{N}_m = \left\{ D_m^{(1)} \leq b_m^{(1)} {I_m^{(1)}}^{-0.5}, \ldots, D_m^{(K_m)} \leq b_m^{(K_m)} {I_m^{(K_m)}}^{-0.5} \right\}, \quad (2.8)$$

$\mathbf{D}_m$ has a truncated multivariate normal distribution (Cohen, 1991) and $D_m = D_m^*$. The distribution of D_m conditional on non-truncation is therefore the distribution of $D_m^* | \, \mathscr{N}_m$, which is the K_m+1 margin of the distribution of $\mathbf{D}_m | \, \mathscr{N}_m$. In the following subsections we make use of the moments of this distribution to gain insight into the properties of the final treatment difference conditional on non-truncation of an RCT. This can be undertaken theoretically for the case of a single interim analysis with $K_m = 1$, while for $K_m > 1$ it can be undertaken by simulation using (2.7) and (2.8) together with a truncated multivariate normal simulator (Wilhelm and Manjunath, 2012).

2.3.1 Estimation bias

Initially we suppose that study m has a single interim analysis, that is $K_m = 1$. Let $t_m = I_m^{(1)}/I_m^*$ be the fraction of the total planned information accrued at the interim analysis and let

$\delta_m = \delta\sqrt{I_m^*}$ denote the standardised treatment effect size. Then for non-zero δ, the relative bias of the observed treatment effect D_m^*, conditional on non-truncation, can be expressed in terms of t_m and δ_m. In particular, using details provided in the web-based supplementary materials and the notation $\Lambda(x) = \phi(x)/\Phi(x)$, where ϕ and Φ are the standard normal density and distribution functions, the relative bias is

$$B\big(\delta_m, t_m\big) = \frac{\delta - E\big(D_m^* \big| \mathscr{N}_m\big)}{\delta} = \Lambda\left(b_m^{(1)} - \delta_m\sqrt{t_m}\right)\frac{\sqrt{t_m}}{\delta_m}. \tag{2.9}$$

Since the relative bias B is non-negative for all t_m and δ_m, the first point to observe from (2.9) is that conditioning on non-truncation leads to underestimation of the treatment effect. This underestimation bias can be computed as a function of the power of the study and the information fraction at the time of the interim analysis, using the fact that $\delta_m = \Phi^{-1}(1-\alpha_m) + \Phi^{-1}(1-\beta_m)$, where $1-\beta_m$ is the power to detect the true treatment effect δ. Thus, the form of the relative bias (2.9) is very general in that it does not depend on the absolute magnitude of the treatment effect δ. Panel A of Figure 2.1 provides computations of the underestimation bias for a range of assumptions using the O'Brien-Fleming stopping boundary. It can be seen that the magnitude of the bias can be practically important, and is generally in the range of $10\% - 20\%$ for adequately powered studies where the interim analysis is conducted near the halfway point of the study. The underestimation bias from conditioning on non-truncation is larger when the interim analysis is conducted later in the study and exceeds 20% for adequately powered studies with an interim analysis in the final third of the study.

The above theoretical calculations are for sequential monitoring with a single interim analysis. As noted above, generalisation of the results to multiple interim analyses can be undertaken using a truncated multivariate normal simulator (Wilhelm and Manjunath, 2012). Such simulations have been carried out for up to five equally spaced interim analyses under various assumptions about the power of the study, and are summarised in Table 2.1. For a single interim analysis, there was close agreement between the theoretical and simulated bi-

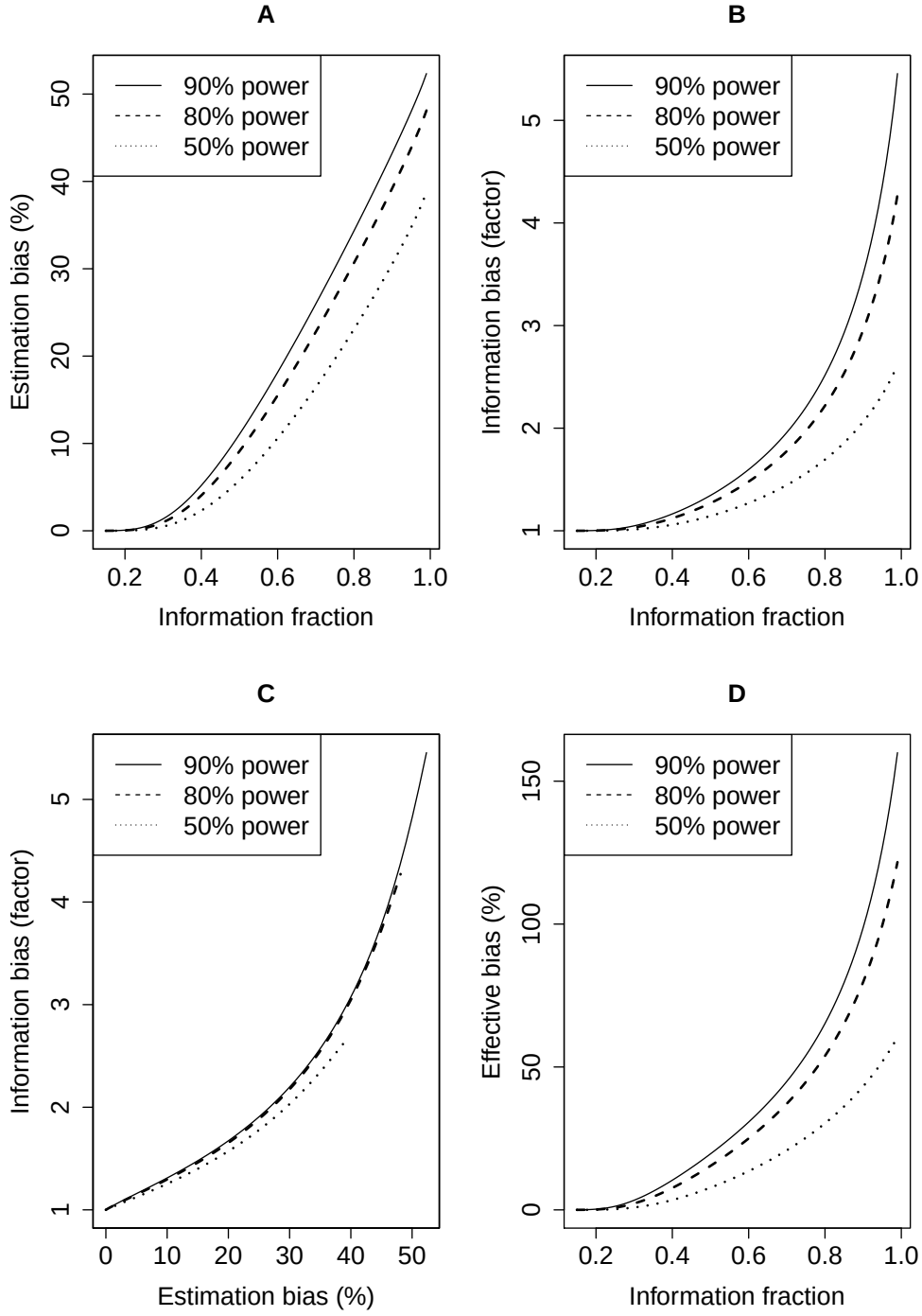

Figure 2.1: Estimation bias (percent underestimation), information bias (inflation factor) and effective estimation bias (percent underestimation), conditional on non-truncation, plotted against the information fraction at which a single interim analysis is conducted with O'Brien-Fleming boundary (Panels A, B and D), and information bias plotted against estimation bias (Panel C). Results are displayed for various values of the power to detect the true treatment effect δ.

ases. Three boundaries were considered: Haybittle-Peto, O'Brien-Fleming and Pocock. Of these, the Haybittle-Peto boundary was the least biased, while the O'Brien-Fleming boundary introduced less bias than the Pocock boundary when $K_m \leq 2$ and was similar to the Pocock boundary when $K_m > 2$. These simulations show that the underestimation due to conditioning on non-truncation increases with more frequent interim monitoring. For example, with the O'Brien-Fleming boundary, the bias is on the order of 30% when there are three interim analyses.

Table 2.1: Estimation bias expressed as percent underestimation, and information bias expressed as an inflation factor, conditional on non-truncation of group sequential studies with K_m equally spaced interim analyses. Results are based on theoretical and simulated bias, with 90% power to detect the true treatment effect δ and one-sided significance level 2.5%.

Boundary	K_m	Theoretical bias		Simulated bias	
		Estimation	Information	Estimation	Information
O'Brien-Fleming	1	11.11	1.35	11.03	1.34
	2	–	–	23.53	1.82
	3	–	–	30.22	2.21
	4	–	–	34.61	2.52
	5	–	–	37.26	2.80
Haybittle-Peto	1	8.05	1.27	8.06	1.26
	2	–	–	14.75	1.55
	3	–	–	19.29	1.81
	4	–	–	22.51	2.03
	5	–	–	25.15	2.20
Pocock	1	19.70	1.52	19.67	1.52
	2	–	–	28.18	1.96
	3	–	–	32.44	2.29
	4	–	–	35.12	2.54
	5	–	–	36.74	2.73

2.3.2 Information bias

The bias resulting from conditioning on non-truncation is not limited to estimation bias. As discussed in Section 2.2.1, meta-analyses typically weight the study-specific treatment effects by a measure of the statistical information in each study. In this section we see that these information weights are also subject to bias when conditioning on non-truncation. This bias can be quantified by comparing the unconditional inverse variance measure of statistical

information I_m^*, with the actual Fisher information based on the truncated normal distribution that accommodates the conditional nature of the sampling. The difference between these unconditional and conditional versions of information is the information bias.

For a theoretical discussion we return to the case of a single interim analysis with known variance. The treatment effects at the interim and final analyses, $\left(D_m^{(1)}, D_m^*\right)$, are distributed bivariate normal according to (2.7) with $K_m = 1$. Let $f_m(d_1, d_2; \delta)$ be the density function corresponding to this bivariate distribution, and let $F_m^{(1)}(d_1; \delta)$ be the marginal distribution function for $D_m^{(1)}$. It follows using (2.8) that the conditional bivariate density function of $\left(D_m^{(1)}, D_m^*\right)$, given non-truncation, is

$$f_m\left(d_1, d_2; \delta \middle| \mathscr{N}_m\right) = \frac{f_m\left(d_1, d_2; \delta\right)}{F_m^{(1)}\left(c_m^{(1)}; \delta\right)} \qquad d_1 \in (-\infty, c_m^{(1)}] \quad d_2 \in (\infty, \infty)$$

where $c_m^{(1)} = b_m^{(1)} {I_m^{(1)}}^{-0.5}$. Based on this density function the Fisher information $\mathscr{I}\left(\delta \middle| \mathscr{N}_m\right)$ can be determined in the standard way, as the negative expected second derivative of

$$\log f_m\left(D_m^{(1)}, D_m^*; \delta \middle| \mathscr{N}_m\right).$$

The form of $\mathscr{I}\left(\delta \middle| \mathscr{N}_m\right)$, which depends on Λ in a complicated manner, is provided in the web-based supporting materials and satisfies

$$I_m^* = \mathscr{I}\left(\delta \middle| \mathscr{N}_m\right) \mathscr{F}\left(\delta_m, t_m\right) \qquad \text{where} \qquad \mathscr{F}(\delta_m, t_m) \geq 1. \tag{2.10}$$

Equation (2.10) means that there is an inflation factor $\mathscr{F}$ that quantifies the bias in I_m^* as a measure of statistical information from study m, when conditioning on non-truncation. In particular, conditional on non-truncation, it follows that the standard approach to study-specific weighting will lead to overestimation of the statistical information. As with the relative estimation bias specified in (2.9), the extent of this information bias is again a function of the power of the study to detect the true treatment effect δ and the information fraction at the time of the interim analysis.

Panel B of Figure 2.1 presents the information bias and its relationship with the power of the study and the information fraction at the time of the interim analysis. It is seen that the information bias increases with increasing power and increasing information fraction. While it is clear that the information bias can be substantial, the magnitude of the information bias is of less interest when viewed in isolation than when viewed in conjunction with the estimation bias, which we defer until the next subsection.

The simulation studies of estimation bias described in Section 2.3.1 can also be used to study the information bias through simulation. Using properties of the truncated bivariate normal distribution, as explained further in the web-based supporting materials, the unconditional and conditional variance of D_m^* can be related according to

$$\text{var}\left(D_m^*\right) = \text{var}\left(D_m^* \middle| \mathscr{N}_m\right)\mathscr{F}\left(\delta_m, t_m\right). \tag{2.11}$$

Thus, the ratio of the unconditional and conditional simulated variance of D_m^* can be used to study the information inflation factor $\mathscr{F}$ through simulation. This is presented in Table 2.1, where it is again seen that the simulation studies are in close agreement with the theoretical studies for $K_m = 1$. Also shown in Table 2.1 is the simulated variance ratio when there are multiple interim analyses, showing an increase in information bias with increasing number of interim analyses, as was also the case for the estimation bias results presented in Section 2.3.1.

2.3.3 Relationship between estimation and information bias

In the previous two subsections we have seen that conditioning on non-truncation leads to underestimation of the treatment effect and overestimation of the statistical information. In this section we see that these two biases are in fact related, in the sense that a monitoring pattern that leads to greater estimation bias conditional on non-truncation, will also lead to greater information bias. Thus, under a strategy of excluding truncated studies from a meta-analysis, a double whammy occurs in which greater estimation bias is associated with

greater overweighting in the aggregated treatment effect estimate.

The association between the two biases can be investigated by considering a study with given power to detect the true treatment difference at the final analysis, that is, a study with given standardised treatment difference δ_m. With a single interim analysis it can be shown theoretically, as detailed in the web-based supporting materials, that the estimation and information biases are both increasing functions of the information fraction at the interim analysis, that is

$$B\big(\delta_m, u\big) \geq B\big(\delta_m, v\big) \quad \text{and} \quad \mathscr{F}\big(\delta_m, u\big) \geq \mathscr{F}\big(\delta_m, v\big) \quad \text{for} \quad u \geq v. \tag{2.12}$$

This relationship was borne out in our theoretical calculations of Sections 3.1 and 3.2 for a single interim analysis, and was also evident in our simulations of studies with more than one interim analysis. It follows that monitoring patterns which induce greater estimation bias conditional on non-truncation will also lead to greater information bias. This theoretical finding is illustrated in Panel C of Figure 2.1 for the $K_m = 1$ case, where the information bias is seen to increase as the estimation bias increases. Furthermore, the simulation results presented in Table 2.1 show the same pattern when $K_m > 1$.

An implication of this relationship is that greater estimation bias is accompanied by greater overweighting in a meta-analysis restricted to non-truncated studies. This suggests that the combined effect of estimation and information bias will be greater than the estimation bias alone. To assess this, the product of the conditional treatment effect expectation and information $E(D_m^*|\ \mathscr{N}_m)\mathscr{I}(\delta|\ \mathscr{N}_m)$, can be compared with its unconditional counterpart δI_m^*. The relative magnitude of these two quantities measures the combined effective bias of the information weighted treatment effect. This is presented in Panel D of Figure 2.1, where it is seen that the effective bias increases with increasing power. Furthermore, by comparison with Panel A of Figure 2.1, it can be seen that the effective bias exceeds the unweighted estimation bias. This reflects the double whammy of combined estimation bias and information bias in meta-analyses that exclude truncated studies.

2.4 Meta-analysis strategies

The distribution of D_m conditional on non-truncation, as discussed in Section 2.3, illustrates the bias that is introduced by excluding truncated studies from estimation of treatment effects. The usefulness of these results for understanding the extent of bias in meta-analyses is limited to situations in which all studies in a meta-analysis are subject to interim monitoring and the fixed effects meta-analysis model is used. In practice, meta-analyses typically involve a mix of sequential and fixed design studies, and may involve the used of a random effects model. In this section we present a simulation study of meta-analyses conducted under various assumptions about the mix of sequential and fixed designs, using both the fixed and random effects models. The primary goal is to make comparisons between three possible strategies for undertaking the meta-analyses: (i) the *non-truncated strategy* in which only studies that proceed to the final analysis are included in the meta-analysis; (ii) the *non-sequential strategy* in which only studies that were not subject to interim analysis are included in the meta-analysis; and (iii) the *all-study strategy* in which all studies, both truncated and non-truncated, are included in the meta-analysis.

The assumptions of the meta-analysis simulations can be described as follows. For each meta-analysis simulation, data from a normally distributed endpoint with effect size δ/σ were simulated for a total of M studies with either 80% or 90% power. The values considered for δ/σ were 0.1, 0.25 and 0.5, and for M were 4, 12 and 24. Of the M studies in each simulation, 0%, 25%, 50%, 75% or 100% were subject to interim monitoring while the remainder were fixed design studies. The sequential studies had equally spaced interim analyses with $K_m = 1$, 2, 3, 4 or 5, conducted using the O'Brien-Fleming, Haybittle-Peto or Pocock stopping boundaries. Both the fixed and random effects meta-analysis estimates described in Section 2.2.1 were calculated for the three meta-analyses strategies, using sample-based variance estimation for the information weights in (2.2).

2.4.1 Non-truncated strategy

Based on the simulation results, Table 2.2 and Panel A of Figure 2.2 present the relative estimation bias in meta-analyses conducted under the non-truncated strategy, in which any study terminated at an interim analysis is excluded. The results displayed are for meta-analyses with $M = 12$ studies monitored using an O'Brien-Fleming stopping boundary and powered to detect a treatment effect of $\delta/\sigma = 0.25$ with 90% power. These simulation results again show that the meta-analysis estimate is an underestimate of the treatment effect and that the underestimation increases as the number of interim analyses increases. As expected, when the proportion of studies subject to sequential analysis is low the bias from excluding truncated studies is low. When at least half the studies are subject to sequential monitoring the underestimation bias is on the order of $5-15\%$, regardless of whether the fixed or random effects approach was used. When all studies are subject to interim monitoring then the bias can be substantially greater than that range. Consistent results were found for other simulation combinations, as displayed in the web-based supporting materials. Overall, these simulation results show that a strategy of excluding truncated studies from meta-analyses introduces bias into the estimation of treatment effects.

2.4.2 Non-sequential strategy

As the non-truncated strategy leads to estimation bias, an alternative approach would be to exclude all sequentially monitored trials thereby avoiding the issue of estimation bias, since meta-analysis estimates from such an approach are theoretically unbiased. Simulation results that demonstrate this are presented in Table 2.2 and the web-based supporting materials. The drawback of this strategy is that it discards information contained in the sequentially monitored studies. Thus, if a strategy is available that is not subject to estimation bias and uses the information contained in the sequentially monitored studies, then the non-sequential strategy would be expected to have reduced efficiency relative to such a strategy. This efficiency reduction is discussed in more detail in the next subsection.

Table 2.2: Simulated estimation bias (percent underestimation) for non-truncated, non-sequential and all-study strategies in meta-analyses of $M = 12$ studies, each with 90% power for an effect size of $\delta/\sigma = 0.25$ at a one-sided significance level of 2.5%. All studies used the O'Brien-Fleming stopping boundary with K_m equally space interim analyses. Values in parentheses for the non-sequential strategy are the efficiencies (percent) relative to the all-study strategy.

K_m	Sequential (%)	Non-truncated (%)	Estimation bias					
			Non-truncated strategy		Non-sequential strategy		All-study strategy	
			Fixed	Random	Fixed	Random	Fixed	Random
1	0	100	0.10	0.09	0.10 (100)	0.09 (100)	0.10	0.09
2	0	100	-0.21	-0.21	-0.21 (100)	-0.21 (100)	-0.21	-0.21
3	0	100	-0.09	-0.10	-0.09 (100)	-0.10 (100)	-0.09	-0.10
4	0	100	0.21	0.22	0.21 (100)	0.22 (100)	0.21	0.22
5	0	100	1.38	1.38	1.38 (100)	1.38 (100)	1.38	1.38
1	25	92	-1.89	-1.88	0.59 (81)	0.61 (82)	0.15	0.34
2	25	86	-2.55	-2.55	0.00 (84)	0.00 (85)	0.14	0.32
3	25	82	-2.64	-2.64	-0.02 (86)	-0.02 (86)	0.16	0.40
4	25	80	-1.63	-1.63	0.37 (91)	0.38 (92)	0.94	1.11
5	25	81	-2.26	-2.26	0.03 (88)	0.04 (89)	0.21	0.46
1	50	84	-3.98	-3.99	0.66 (47)	0.64 (47)	0.86	1.25
2	50	71	-7.51	-7.52	-0.76 (62)	-0.76 (63)	-0.53	-0.06
3	50	65	-6.80	-6.80	-0.53 (59)	-0.52 (61)	0.18	0.72
4	50	63	-7.19	-7.18	-0.91 (59)	-0.89 (59)	-0.56	0.00
5	50	60	-7.31	-7.31	-1.69 (53)	-1.69 (54)	-0.74	-0.14
1	75	76	-6.31	-6.32	1.43 (30)	1.43 (31)	1.53	2.34
2	75	56	-12.26	-12.23	0.02 (38)	0.07 (39)	0.18	0.88
3	75	47	-12.96	-12.96	-0.12 (30)	-0.12 (31)	1.08	1.94
4	75	43	-12.30	-12.27	0.26 (35)	0.29 (37)	1.20	2.19
5	75	41	-12.44	-12.45	0.22 (33)	0.20 (35)	1.46	2.42
1	100	69	-11.42	-11.40	—	—	0.34	1.41
2	100	41	-23.16	-23.08	—	—	1.10	2.35
3	100	30	-29.48	-29.63	—	—	1.05	2.39
4	100	23	-35.39	-34.99	—	—	1.14	2.84
5	100	23	-36.98	-37.13	—	—	0.10	1.69

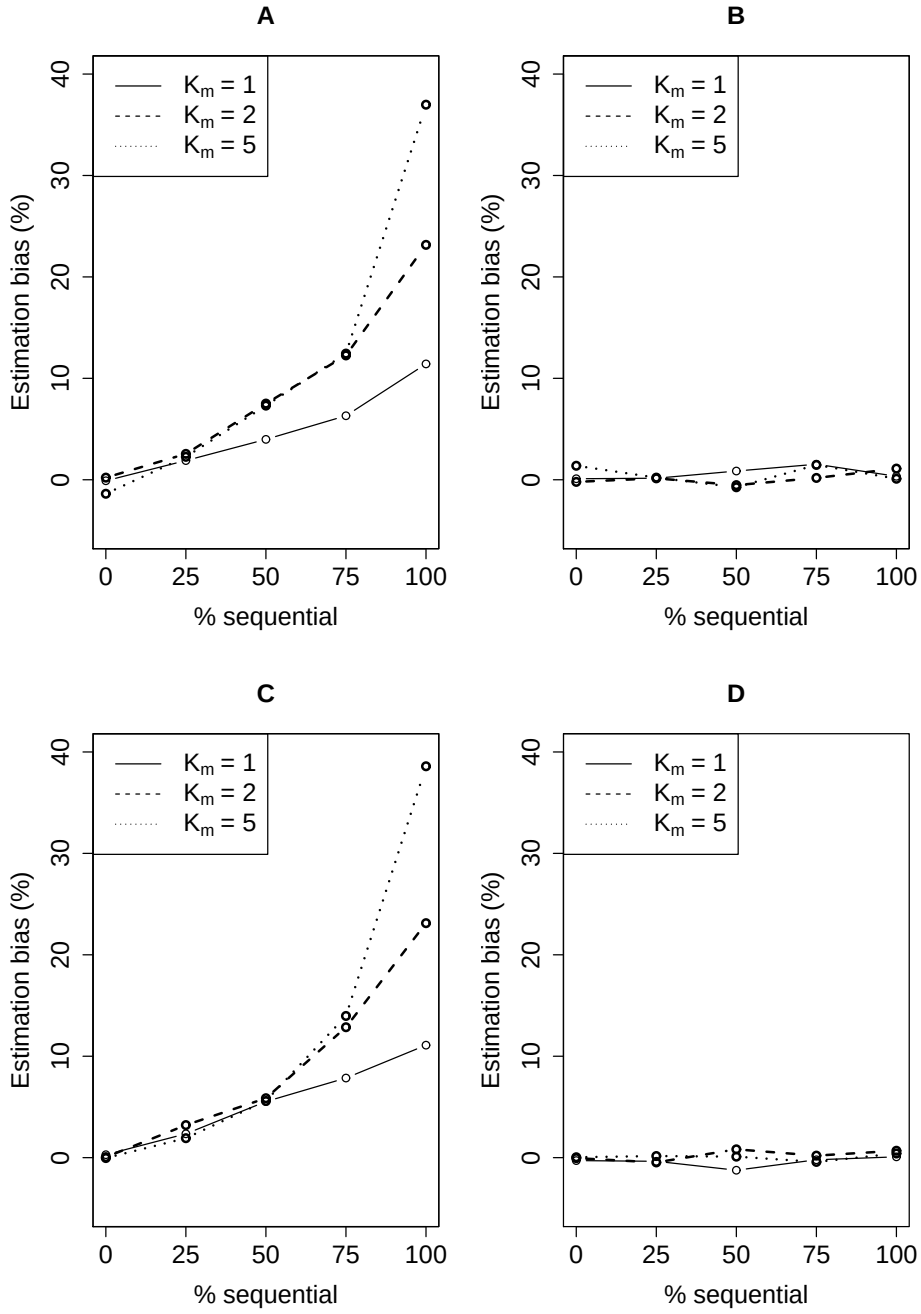

Figure 2.2: Simulation results for meta-analyses of $M = 12$ studies with various percentages of studies subject to sequential monitoring using the O'Brien-Fleming boundary with various numbers of interim analyses (K_m). In each panel, estimation bias (percent underestimation) is plotted against the percent of studies with sequential monitoring, using the non-sequential strategy with mean differences (Panel A) and risk differences (Panel C), or the all-study strategy with mean differences (Panel B) and risk differences (Panel D). Results are displayed for studies with 90% power to detect an effect size of $\delta/\sigma = 0.25$ for mean differences, and $p_1 = 0.1$ versus $p_2 = 0.25$ for risk differences.

2.4.3 All-study strategy

The final strategy is to include all studies in the meta-analysis. The simulation results for this approach are also presented in Table 2.2, the web-based supporting materials and graphically in Panel B of Figure 2.2. The results suggest that the treatment effect estimates have no notable bias when all studies are included in the meta-analysis, irrespective of the number of interim analyses and the proportion of studies that are subject to interim monitoring. There is some suggestion that the random effects estimates may have slightly larger bias than the fixed effects estimates, however, in both cases any bias is extremely small. The efficiency gain made by the all-study strategy over the non-sequential strategy is also presented in Table 2.2 and supports the argument in favour of including all studies. For example, the efficiency of the non-sequential strategy was on the order of 80% when 25% of the studies were subject to interim monitoring, and was on the order of 30% when the proportion subject to interim monitoring increased to 75%. Thus, given that the all-study strategy is relatively unbiased, the recommended approach to evidence synthesis in meta-analysis would be to include all studies, as this approach results in greater efficiency than the non-sequential strategy and does not possess the bias of the non-truncated strategy.

2.4.4 Alternative effect measures

Results similar to those discussed above also apply to effect measures related to binary outcomes. Assuming the outcome variable X_{ij} introduced in Section 2.2.1 is binary (0/1) with $E(X_{ij}) = p_j$ where $p_2 = p_1 + \delta$ and $\text{var}(X_{ij}) = p_j(1-p_j)$, then large sample normality can be used to make use of the theoretical results from Section 2.3. In particular, the risk difference $D = \overline{X}_2 - \overline{X}_1$ is normally distributed with expectation $\delta = p_2 - p_1$ and $n\text{var}(D) = \nu^2 = p_1(1-p_1)/\gamma_1 + p_2(1-p_2)/\gamma_2$. Likewise, the log relative risk $D = \log\left(\overline{X}_2/\overline{X}_1\right)$ is normally distributed with expectation $\delta = \log\left(p_2/p_1\right)$ and $n\text{var}(D) = \nu^2 = (1-p_1)/p_1\gamma_1 + (1-p_2)/p_2\gamma_2$. With these modified forms of ν^2, equation (2.9) can be used in the same way as in Section 2.3 for normally distributed endpoints to produce theoretical results comparable to Figure 2.1. We make use of this approach in the next section. For

now, we limit our presentation to the corresponding results from simulation studies, which are presented for risk differences in Figure 2.2, Panels C and D. These simulations were conducted in the same way as for normally distributed endpoints and were based on $M = 12$ trials with O'Brien-Fleming boundaries and 90% power to detect a risk difference of 0.15 ($p_1 = 0.1$ and $p_2 = 0.25$). Similar results arose from other combinations. It can be seen that the conclusions drawn in the previous sections also hold when a risk difference is used to measure the treatment effect. In particular, the underestimation bias arising from the non-truncated strategy can be substantial and increases as the proportion of studies subjected to interim monitoring increases or the number of interim analyses per study increases. On the other hand, no systematic bias is present in the aggregated estimate based on the all-study strategy.

2.5 Comparison of truncated and non-truncated studies

Previous research on potential biases resulting from early stopping of RCTs has compared meta-analyses of truncated studies with meta-analyses of non-truncated studies addressing the same research question (see Bassler et al. (2010) and references therein). This approach has identified larger estimated treatment effects in meta-analyses that include truncated studies. This empirical observation has been interpreted as demonstrating that inclusion of truncated studies in meta-analyses leads to overestimation of treatment effects, a conclusion that is at odds with the results presented earlier in this paper. In this section we demonstrate that it is flawed to compare truncated and non-truncated studies in order to assess whether truncated studies are biased. We show that differences between truncated and non-truncated studies are expected purely due to conditioning mechanisms, and are not reflective of any inherent bias in meta-analyses that include truncated studies. Thus, the findings reported earlier in this paper are not at odds with the empirical observations published in previous research, but our results do cast doubt on the conclusions drawn from these empirical observations.

2.5.1 Theoretical comparisons

We begin with a theoretical study of the expected difference between treatment effects from truncated studies and those from non-truncated studies. Assuming $K_m = 1$, the expected treatment effect conditional on truncation is $E\big(D_m^{(1)} | \overline{\mathscr{N}}_m\big)$, where $\overline{\mathscr{N}}_m$ is the complement of the event $\mathscr{N}_m$, that is, $\overline{\mathscr{N}}_m$ is the event that the study stops at the interim analysis. Using details provided in the web-based supporting materials, and the notation $\Psi(x) = \phi(x)/(1 - \Phi(x))$, the analogue of (2.9) for the truncated situation is

$$\overline{B}\big(\delta_m, t_m\big) = \frac{\delta - E\big(D_m^{(1)} | \overline{\mathscr{N}}_m\big)}{\delta} = -\Psi\left(b_m^{(1)} - \delta_m\sqrt{t_m}\right)\frac{1}{\delta_m\sqrt{t_m}}. \tag{2.13}$$

Equation (2.13) demonstrates that the observed treatment effect conditional on truncation is an overestimate of δ, since the relative bias $\overline{B}$ is negative for all t_m and δ_m. As with the underestimation bias in the non-truncated case, the overestimation bias in (2.13) is a function of the power of the study and the information fraction at the time of the interim analysis. This relationship is presented in Panel A of Figure 2.3 using the O'Brien-Fleming stopping boundary and shows that for studies where the interim analysis is conducted at 50% information fraction, the observed treatment effect conditional on truncation is expected to be more than 50% greater than the true treatment effect δ.

Of more interest is the difference between the expected treatment effect conditional on truncation and the expected treatment effect conditional on non-truncation, since this difference has previously been used to infer bias in truncated studies. Using (2.9) and (2.13) this difference is $\mathscr{B}\big(\delta_m, t_m\big) = B\big(\delta_m, t_m\big) - \overline{B}\big(\delta_m, t_m\big)$, which is a combination of the biases presented in Panel A of Figures 2.1 and 2.3. Plots of $\mathscr{B}\big(\delta_m, t_m\big)$ are presented in Panel B of Figure 2.3. The main points to note are that differences in the treatment effect estimates between truncated and non-truncated studies are to be expected due to the conditioning mechanisms and that these differences can be quite substantial. Such expected differences are simply analogous to the expected differences that exist between the more extreme and the less extreme observations in a sample, and do not reflect any inherent bias resulting from stopping

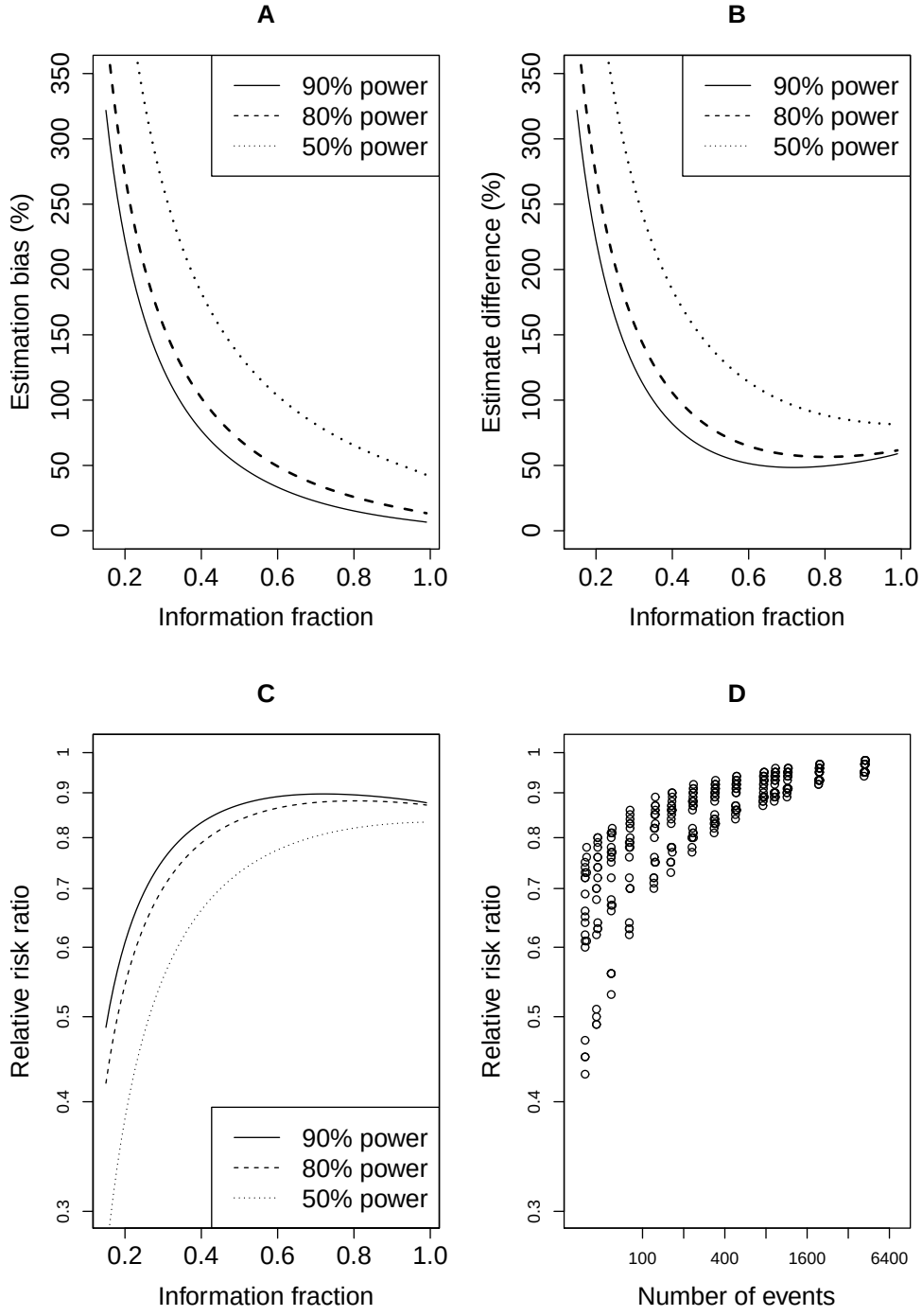

Figure 2.3: Estimation bias (percent overestimation) conditional on truncation (Panel A), and differences in treatment effects from truncated and non-truncated studies, using either percent difference between mean differences (Panel B) or the ratio of relative risks (Panel C). Results are displayed for various values of the power to detect the true treatment effect δ. Also displayed are simulations of the relationship between the number of outcome events and the ratio of relative risks from truncated and non-truncated studies, under a range of assumptions described in Section 2.5.2 (Panel D).

studies early.

While the above comments relate to treatment effects measured using mean differences, they are also relevant to situations in which other measures of treatment effect are used. A ratio of relative risks, comparing the relative risk conditional on truncation with the relative risk conditional on non-truncation, has been used previously as a measure of the difference in treatment effects (Bassler et al., 2010). As discussed in Section 2.4.4, the estimation biases (2.9) and (2.13) can be used with a modified v^2 to study the expected log relative risk under truncation and non-truncation, respectively. Thus,

$$\exp\{\delta \mathscr{B}(\delta_m, t_m)\} = \frac{\exp\{E(D_m^{(1)} | \overline{\mathscr{N}_m})\}}{\exp\{E(D_m^* | \mathscr{N}_m)\}} \tag{2.14}$$

can be used to quantify the ratio of the relative risks from truncated and non-truncated studies, where $D_m^{(1)}$ and D_m^* are now log relative risks rather than mean differences. Note that unlike the estimation biases (2.9) and (2.13), the ratio of relative risks (2.14) is not independent of the true treatment effect δ.

Panel C of Figure 2.3 presents computations of (2.14) using the O'Brien-Fleming stopping boundary and assuming a true relative risk of 0.8. A similar conclusion to that of Panel B of Figure 2.3 can be drawn, noting that smaller values of the relative risk ratio correspond to larger differences between truncated and non-truncated studies. In particular, it can be seen that relative risks from truncated studies can be expected to be at least 10% less than those of non-truncated studies, and often much more than 10%. Similar differences are also evident for other values of the true relative risk. As with the results on treatment effects measured using mean differences, these results for relative risks are simply a consequence of comparing more extreme observations (truncated studies) with less extreme observations (non-truncated studies) and do not reflect any inherent bias in studies stopped early.

2.5.2 Empirical comparisons

The theoretical calculations for a single interim analysis given in the previous subsection show that observed treatment effects conditional on truncation are expected to be larger than those conditional on non-truncation. In this subsection we examine these differences for more general situations and make comparisons with empirical analyses from the literature.

A comprehensive empirical comparison of relative risk estimates from truncated studies with those from non-truncated studies which addressed the same clinical questions was conducted as part of the STOPIT-2 study (Bassler et al., 2010). In all, 515 RCTs were identified addressing 63 clinical questions. Of these, 91 were truncated RCTs with a majority of the clinical questions (43/63) being addressed by a single truncated RCT, and a maximum of four truncated studies per question. The matching non-truncated RCTs addressing the 63 questions ranged from one to 38 studies per question. The evidence from these studies was synthesised using random effects inverse variance meta-analysis and pooled estimates of relative risks were calculated. The ratios of these relative risks were then examined and related to other study characteristics such as the total number of events observed in the study. The authors observed that truncated RCTs were associated with greater treatment effects than non-truncated RCTs, and used this observation to conclude that truncated RCTs are biased. As a consequence, they cautioned that if systematic reviewers did not take into consideration early stopping for benefit, then meta-analyses were likely to report overestimates of treatment effect. Furthermore, they stated that while it was not possible to know whether the observed treatment effect is biased in any particular study, their findings suggested that estimates from truncated studies were often not close to the truth.

The above empirical observations are equally well explained by the effects of the conditioning mechanisms involved in these comparisons. This was illustrated theoretically in the previous subsection and here we present results of a more general simulation study addressing this issue. A sample of binary random variables as described in Section 2.4.4 were generated, with $M = 12$, K_m ranging from 1 through 5, power of 90%, $p_1 = 0.1$ and p_2

taking on a range of values from 0.110 to 0.275 corresponding to relative risks ranging from 0.36 to 0.91. In each simulation, a proportion of these $M = 12$ studies (25%, 50%, 75% or 100%) was subject to interim analyses using O'Brien-Fleming stopping boundaries and the remainder were simulated as fixed design studies. Both fixed effects and random effects inverse variance meta-analyses were conducted and the estimates of relative risks were obtained for the truncated and non-truncated studies. The ratios of the fixed effects relative risks are presented in Panel D of Figure 2.3. Further information on these simulations, including comparison of the fixed effects and random effects results, is presented in Table 2.3.

Based on these results, the following observations can be made. Firstly, even in the absence of any bias, it is clear that one should expect to see the ratio of relative risks from truncated and non-truncated studies depart substantially from 1, and typically lie in the range $0.5 - 1$. This is highly consistent with the empirical comparisons presented by Bassler et al. (2010). Secondly, our results show that one should also expect to see a relationship between this difference and the total number of events, with greater differences arising for smaller numbers of events. In particular, Panel D of Figure 2.3 is highly consistent with the analogous empirical plot presented as Figure 3 in Bassler et al. (2010), and is indicative not of bias but rather the conditioning involved in separating and comparing the results of truncated and non-truncated studies.

Overall, these results indicate that it should not be surprising that observed treatment effects from truncated studies tend to be greater than those from non-truncated studies. When taken together with the results of Section 2.4 on meta-analyses involving truncated studies, our results suggest that truncated studies do not introduce bias into meta-analyses and that it is flawed to investigate bias by comparing the results of truncated and non-truncated studies.

Table 2.3: Simulated ratios of relative risks (truncated strategy divided by non-truncated strategy) from meta-analyses of $M = 12$ studies, each with 90% power for a range of relative risks with a one-sided significance level of 2.5%. All studies had $p_1 = 0.10$ and used the O'Brien-Fleming stopping boundary with $K_m = 2$ equally spaced interim analyses.

p_2	Relative risk	Sequential (%)	Non-truncated (%)	Relative risk ratio	
				Fixed effects	Random effects
0.110	0.91	25	85	0.969	0.969
0.120	0.83	25	85	0.948	0.947
0.125	0.80	25	85	0.935	0.934
0.150	0.67	25	86	0.875	0.873
0.175	0.57	25	86	0.823	0.822
0.200	0.50	25	86	0.791	0.792
0.225	0.44	25	87	0.767	0.767
0.250	0.40	25	87	0.702	0.704
0.275	0.36	25	88	0.641	0.638
0.110	0.91	50	70	0.969	0.969
0.120	0.83	50	71	0.941	0.940
0.125	0.80	50	70	0.924	0.922
0.150	0.67	50	71	0.877	0.873
0.175	0.57	50	72	0.821	0.820
0.200	0.50	50	72	0.783	0.783
0.225	0.44	50	72	0.762	0.764
0.250	0.40	50	74	0.699	0.700
0.275	0.36	50	75	0.663	0.663
0.110	0.91	75	56	0.965	0.965
0.120	0.83	75	55	0.931	0.931
0.125	0.80	75	56	0.914	0.914
0.150	0.67	75	56	0.855	0.854
0.175	0.57	75	58	0.799	0.798
0.200	0.50	75	58	0.763	0.763
0.225	0.44	75	59	0.720	0.722
0.250	0.40	75	62	0.678	0.680
0.275	0.36	75	62	0.649	0.650
0.110	0.91	100	43	0.953	0.953
0.120	0.83	100	42	0.913	0.913
0.125	0.80	100	43	0.895	0.895
0.150	0.67	100	43	0.820	0.820
0.175	0.57	100	42	0.761	0.761
0.200	0.50	100	42	0.716	0.715
0.225	0.44	100	46	0.669	0.669
0.250	0.40	100	47	0.619	0.619
0.275	0.36	100	51	0.596	0.596

2.6 Discussion

There is ongoing debate concerning possible bias in studies stopped early for benefit and particularly the potential effect on meta-analyses. This is bound to cause some systematic reviewers to consider excluding truncated studies from evidence synthesis. In this paper we have seen that such a strategy is biased, in that it leads to underestimation of treatment effects. Furthermore we have seen that the standard approach of aggregating the results of all studies is essentially unbiased.

We investigated the effect of excluding truncated studies by examining the properties of sequentially monitored studies conditional on non-truncation, and by studying the expected differences between truncated and non-truncated studies. Potential biases were quantified using both theoretical investigations and simulations, and comparisons were made with previously published empirical observations. We found that excluding truncated studies leads to underestimation of treatment effects and overestimation of statistical information. For meta-analyses using inverse-variance weighting, these biases lead to a double whammy in which greater estimation bias is accompanied by greater overweighting. In meta-analysis simulations we found that the bias resulting from excluding truncated studies can be substantial, whereas inclusion of all studies leads to effectively unbiased estimation. The magnitude of any bias tended to be greater when studies were subjected to more frequent interim analysis and when the proportion of studies with sequential monitoring was greater.

We also examined previous empirical observations that have been used to cast doubt over the validity of early stopping, including larger treatment effects in truncated studies and greater differences between truncated and non-truncated studies in smaller trials (Montori et al., 2005; Bassler et al., 2010). These observations have received prominence in the published literature and were even noted in the most recent version of the CONSORT guidelines (Chen et al., 2010). Importantly, we found that such empirical observations are consistent with the expected difference that results from comparing a more extreme study to a less extreme study, rather than reflecting any inherent bias in early stopping.

We have focused here on examining the effect of excluding truncated studies from meta-analyses. An alternative response to the argument that truncated studies introduce bias is to abandon early stopping altogether. While this approach certainly leads to unbiased estimation, it is currently not practical given the range of ethical and practical imperatives involved in clinical trials research. Furthermore, based on our results, such an approach is inefficient and unnecessary.

Although studies that stop early for benefit tend to overestimate the true treatment effect, it should not be too surprising that their inclusion in meta-analyses does not lead to biased estimation. While it is true that conditional on truncation, the observed treatment difference overestimates the true effect, we have seen here that conditional on non-truncation, the observed treatment difference underestimates the true effect. Thus, in aggregate, the effects of truncation and non-truncation tend to balance each other to allow unbiased estimation.

In a preliminary presentation of part of this research, we previously recommended that there should be wider reporting of adjusted estimates for sequentially monitored studies (Schou and Marschner, 2011). However, since the standard all-study strategy performed well in the results presented here, such an approach may not be necessary in practice. While a complete study of the issue is beyond the scope of this paper, we did consider adjustment of estimates in the simulation studies reported in Section 2.4, using methods such as the bias adjusted mean estimate, Rao-Blackwell adjusted estimate, and the median unbiased estimate, which are available in *S+SeqTrial* (*S+SeqTrial User's Manual* 2002). We found that such adjustments did not offer any improvement in bias or efficiency, which is perhaps not unexpected given that the simple all-study approach performed well. Nonetheless, there may be scope for future research to examine this issue in more depth.

Hughes et al. (1992) argued that early stopping of RCTs can introduce artificial heterogeneity between studies leading to greater use of random effects meta-analyses. It is thus important to examine the potential for bias in both the random effects and fixed effects settings. Our investigations did this, and overall our results were very similar in the two contexts.

Our investigations of the effect of early stopping are predicated on the use of statistically valid methods to undertake such early stopping. Clearly, if interim analyses are conducted in a statistically invalid fashion then our conclusions may not continue to hold. Valid stopping boundaries for undertaking interim analysis of clinical trials have been available for some decades now and are generally well understood and implemented in practice.

In summary, we conclude that early stopping of clinical trials for apparent benefit is not a substantive source of bias in meta-analyses, whereas exclusion of truncated studies from meta-analyses would introduce bias. Evidence synthesis should be based on results from all studies, both truncated and non-truncated.

2.7 References

Bassler, D., M. Briel, V. M. Montori, M. Lane, P. Glasziou, Q. Zhou, D. Heels-Ansdell, S. D. Walter, and G. H. Guyatt (2010). Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. *Journal of the American Medical Association* **303**: 1180–1187.

Bassler, D., I. Ferreira-Gonzalez, M. Briel, D. J. Cook, P. J. Devereaux, D. Heels-Ansdell, H. Kirpalani, M. O. Meade, V. M. Montori, A. Rozenberg, H. J. Schünemann, and G. H. Guyatt (2007). Systematic reviewers neglect bias that results from trials stopped early for benefit. *Journal of Clinical Epidemiology* **60**: 869–873.

Bassler, D., V. M. Montori, M. Briel, P. Glasziou, Q. Zhou, and G. H. Guyatt (2008). Early stopping of randomized clinical trials for overt efficacy is problematic. *Journal of Clinical Epidemiology* **61**: 241–246.

Berry, S. M., B. P. Carlin, and J. Connor (2010). Bias and trials stopped early for benefit. *Journal of the American Medical Association* **304**: 156.

Cohen, A. C. (1991). *Truncated and Censored Samples: Theory and Applications*. New York, New York: Marcel Dekker.

Freidlin, B. and E. L. Korn (2009). Stopping clinical trials early for benefit: impact on estimation. *Clinical Trials* **6**: 119–125.

Goodman, S. N. (2008). Systematic reviews are not biased by results from trials stopped early for benefit. *Journal of Clinical Epidemiology* **61**: 95–96.

Hughes, M. D., L. S. Freedman, and S. J. Pocock (1992). The impact of stopping rules on heterogeneity of results in overviews of clinical trials. *Biometrics* **48**: 41–53.

Hughes, M. D. and S. J. Pocock (1988). Stopping rules and estimation problems in clinical trials. *Statistics in Medicine* **7**: 1231–1242.

Jennison, C. and B. W. Turnbull (2000). *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, Florida: Chapman & Hall/CRC.

Montori, V. M. et al. (2005). Randomized trials stopped early for benefit: a systematic review. *Journal of the American Medical Association* **294**: 2203–2209.

Pocock, S. J. and M. D. Hughes (1989). Practical problems in interim analyses with particular regard to estimation. *Controlled Clinical Trials* **10**: 209–221.

Schou, I. M. and I. C. Marschner (2011). Biases in clinical trials with sequential monitoring. *Trials* **12**.Suppl 1: A50.

S+SeqTrial User's Manual (2002). Insightful Corporation. Seattle, Washington.

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

Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials*. Second. Chichester, UK: Wiley.

Wilhelm, S. and B. G. Manjunath (2012). *tmvtnorm: Truncated multivariate normal and t distributions*. R package version 1.4-4.

2.A Appendix: web-based supporting materials for "Meta-analysis of clinical trials with early stopping: an investigation of potential bias"

Summary

In this supporting manuscript we provide proofs of various theoretical results that were referred to in the main paper. These proofs are presented in Sections 2.A.1 through 2.A.3. In Section 2.A.4, we provide additional simulation results for the estimation bias associated with various meta-analysis strategies, to supplement the results presented in Table 2.2 of the main paper.

2.A.1 Proofs of (2.9), (2.11) and (2.13)

Muthén (1990) generalised results of Rosenbaum (1961) by deriving the expectation and variance of each component of the bivariate standard normal random variable (Z_1, Z_2) subject to general upper and lower truncation, $b_1 \leq Z_1 \leq a_1$ and $b_2 \leq Z_2 \leq a_2$. By defining $Z_1 = \left(D_m^{(1)} - \delta\right)/\sqrt{I_m^{(1)}}$ and $Z_2 = \left(D_m^* - \delta\right)/\sqrt{I_m^*}$, it follows from (2.7) with $K_m = 1$ that (Z_1, Z_2) is standard bivariate normal with covariance $\rho = \sqrt{I_m^{(1)}/I_m^*}$. Thus,

$$E\left(D_m^* | \mathscr{N}_m\right) = \delta - \Lambda\left(b_m^{(1)} - \delta\sqrt{I_m^{(1)}}\right)\frac{\sqrt{I_m^{(1)}}}{I_m^*} \tag{2.A.1}$$

follows from the results of Muthén (1990) using the form of the expectation of Z_2 when the random variable (Z_1, Z_2) is subject only to truncation from above in Z_1. In particular, making the substitutions $a_1 = b_m^{(1)} - \delta\sqrt{I_m^{(1)}}$, $a_2 = \infty$, $b_1 = b_2 = -\infty$, $\rho = \sqrt{I_m^{(1)}/I_m^*}$ and $i = 2$ in equation (5) of Muthén (1990), the expectation of Z_2 is obtained, from which the

expectation of D_m^* in (2.A.1) follows. Equation (2.9) follows immediately from (2.A.1). Making the same substitutions in equation (10) of Muthén (1990) leads to equation (2.11). Equation (2.13) can be obtained by first noting that

$$E\left(D_m^{(1)} \middle| \overline{\mathscr{N}}_m\right) = \delta + \Psi\left(b_m^{(1)} - \delta\sqrt{I_m^{(1)}}\right)\frac{1}{\sqrt{I_m^{(1)}}}$$

using the expectation of the univariate standard normal distribution truncated from below at the value $b_m^{(1)} - \delta\sqrt{I_m^{(1)}}$; see Cohen (1991).

2.A.2 Proof of (2.10)

The information conditional on non-truncation can be expressed as

$$\begin{aligned}
\mathscr{I}\left(\delta \middle| \mathscr{N}_m\right) &= -E\left\{\frac{d^2}{d\delta^2}\log f_m\left(D_m^{(1)}, D_m^*; \delta \middle| \mathscr{N}_m\right)\right\} \\
&= -E\left\{\frac{d^2}{d\delta^2}\left[\log f_m\left(D_m^{(1)}, D_m^*; \delta\right) - \log F_m^{(1)}\left(c_m^{(1)}; \delta\right)\right]\right\} \\
&= -E\left\{\frac{d}{d\delta}\left[I_m^*(D_m^* - \delta) + \Lambda\left(b_m^{(1)} - \delta\sqrt{I_m^{(1)}}\right)\sqrt{I_m^{(1)}}\right]\right\} \\
&= I_m^* + I_m^{(1)}\Lambda'\left(b_m^{(1)} - \delta\sqrt{I_m^{(1)}}\right).
\end{aligned}$$

Using the fact that the derivative of the standard normal density function is $\phi'(x) = -x\phi(x)$, it follows that the derivative of $\Lambda(x)$ is

$$\Lambda'(x) = -\Lambda(x)\{x + \Lambda(x)\}. \tag{2.A.2}$$

Substituting equation (2.A.2) into $\mathscr{I}\left(\delta \middle| \mathscr{N}_m\right)$ above yields

$$\begin{aligned}
&\mathscr{I}\left(\delta \middle| \mathscr{N}_m\right) = \\
&I_m^* - I_m^{(1)}\Lambda\left(b_m^{(1)} - \delta\sqrt{I_m^{(1)}}\right)\left\{\left(b_m^{(1)} - \delta\sqrt{I_m^{(1)}}\right) + \Lambda\left(b_m^{(1)} - \delta\sqrt{I_m^{(1)}}\right)\right\}
\end{aligned}$$

which yields the first part of equation (2.10), where

$$\mathscr{F}\left(\delta_m, t_m\right)^{-1} = \\ 1 - t_m \Lambda\left(b_m^{(1)} - \delta_m\sqrt{t_m}\right)\left[\left(b_m^{(1)} - \delta_m\sqrt{t_m}\right) + \Lambda\left(b_m^{(1)} - \delta_m\sqrt{t_m}\right)\right]. \tag{2.A.3}$$

Using (2.A.2), equation (2.A.3) can be written as

$$\mathscr{F}(\delta_m, t_m) = \left\{1 + t_m \Lambda'\left(b_m^{(1)} - \delta_m\sqrt{t_m}\right)\right\}^{-1}. \tag{2.A.4}$$

Since $0 \leq t_m \leq 1$, the second part of equation (2.10) follows if it can be shown that

$$-1 < \Lambda'(x) < 0 \qquad \text{for all } x \in \Re. \tag{2.A.5}$$

Since $\Lambda(x) > 0$ for all $x \in \Re$, $\Lambda'(x) < 0$ follows from (2.A.2) if it can be shown

$$x + \Lambda(x) > 0 \qquad \text{for all } x \in \Re. \tag{2.A.6}$$

Equation (2.A.6) obviously holds for $x \geq 0$. To prove (2.A.6) for $x < 0$, bounds on the tail probability $Q(u) = 1 - \Phi(u)$ can be used. In particular, for $u > 0$, $Q(u) < \phi(u)/u$ (Durrett, 2010). Using $Q(u) = \Phi(-u)$ and $\phi(u) = \phi(-u)$, it follows that $\Phi(-u) < \phi(-u)/u$, or equivalently that $\Lambda(-u) - u > 0$, for $u > 0$. Making the change of variable $x = -u$ therefore leads to (2.A.6) for $x < 0$. This completes the proof of (2.A.6) and hence the right hand inequality in (2.A.5), $\Lambda'(x) < 0$ for all $x \in \Re$.

To prove the left-hand inequality in (2.A.5), $-1 < \Lambda'(x)$ for all $x \in \Re$, we make use of the convexity of Λ and the tail probability bound $Q(u) > \phi(u)u/(1+u^2)$ for $u > 0$ (Pechtl, 1998; Durrett, 2010). These two properties imply, respectively, $\Lambda'(x) > \lim_{u\to-\infty}\Lambda'(u)$ for $x \in \Re$ and $Q(-x) > -\phi(-x)x/(1+x^2)$ for $x < 0$. Thus, using $Q(-x) = \Phi(x)$, $\phi(-x) = \phi(x)$ and the fact that $-x$ is positive, it follows that $\Lambda(x) < -x^{-1} - x$ for $x < 0$. Substituting this into (2.A.2) yields $\Lambda'(x) > -(-x^{-1} - x)(-x^{-1} - x + x) = -1 - x^{-2}$. Therefore, $\lim_{u\to-\infty}\Lambda'(u) >$

$\lim_{u\to-\infty}(-1-u^{-2})=-1$ and consequently $\Lambda'(x)>-1$ for all $x\in\Re$. This completes the proof of (2.A.2) and hence the second part of (2.10).

2.A.3 Proof of (2.12)

The stopping boundary $b_m^{(1)}=b_m^{(1)}(t_m)$ is a non-increasing function of t_m for all stopping rules considered in this paper and indeed any other sensible rule. It follows that $b_m^{(1)}-\delta_m\sqrt{t_m}$ is a decreasing function of t_m for given δ_m. Therefore, since (2.A.3) implies Λ is a decreasing function, equation (2.9) implies $B(\delta_m,t_m)$ is an increasing function of t_m for given δ_m. Similarly, (2.A.2) and the convexity of Λ (Pechtl, 1998), imply that $\mathscr{F}(\delta_m,t_m)$ is an increasing function of t_m for given δ_m. This completes the proof of (2.12).

2.A.4 Additional estimation bias simulations

Table 2.2 of the main paper provided simulation results for the estimation bias associated with various meta-analysis strategies. Additional versions of Table 2.2 are provided below using various alternative parameter combinations as detailed in the table captions.

Table 2.A.1: Simulated estimation bias (percent underestimation) for non-truncated, non-sequential and all-study strategies in meta-analyses of $M = 4$ studies, each with 90% power for an effect size of $\delta/\sigma = 0.25$ at a one-sided significance level of 2.5%. All studies used the O'Brien-Fleming stopping boundary with K_m equally space interim analyses. Values in parentheses for the non-sequential strategy are the efficiencies (percent) relative to the all-study strategy. S and N are the percent of studies with sequential monitoring and non-truncation, respectively.

K_m	S	N	Estimation bias					
			Non-truncated strategy		Non-sequential strategy		All-study strategy	
			Fixed	Random	Fixed	Random	Fixed	Random
1	0	100	1.51	1.47	1.51 (100)	1.47 (100)	1.51	1.47
2	0	100	-0.51	-0.52	-0.51 (100)	-0.52 (100)	-0.51	-0.52
3	0	100	0.66	0.66	0.66 (100)	0.66 (100)	0.66	0.66
4	0	100	0.05	0.05	0.05 (100)	0.05 (100)	0.05	0.05
5	0	100	-0.44	-0.43	-0.44 (100)	-0.43 (100)	-0.44	-0.43
1	25	94	-2.16	-2.12	-1.07 (73)	-1.02 (73)	-0.64	-0.44
2	25	84	-1.50	-1.50	0.23 (74)	0.24 (74)	0.81	0.99
3	25	80	-1.59	-1.59	-0.07 (71)	-0.07 (70)	0.08	0.25
4	25	80	-3.28	-3.26	-2.14 (69)	-2.11 (68)	-1.79	-1.68
5	25	78	-0.90	-0.91	0.06 (71)	0.05 (70)	0.62	0.80
1	50	84	-3.92	-3.95	-0.03 (60)	-0.01 (61)	0.78	1.35
2	50	70	-5.02	-5.01	0.67 (55)	0.67 (57)	0.98	1.88
3	50	67	-4.66	-4.66	1.28 (60)	1.26 (61)	1.47	2.25
4	50	63	-4.80	-4.80	1.31 (60)	1.29 (61)	1.44	2.43
5	50	62	-3.21	-3.21	2.75 (61)	2.77 (64)	1.85	2.82
1	75	76	-8.31	-8.16	0.85 (26)	–	0.71	2.19
2	75	57	-10.61	-13.42	0.07 (27)	–	1.91	3.70
3	75	48	-11.78	-18.88	-1.25 (30)	–	0.23	1.84
4	75	45	-10.92	-17.61	2.04 (34)	–	1.45	3.05
5	75	39	-9.01	-19.28	-0.03 (32)	–	2.31	3.85
1	100	70	-11.35	-11.46	–	–	0.52	1.91
2	100	41	-22.98	-22.83	–	–	1.36	3.13
3	100	32	-27.92	-27.96	–	–	2.15	4.09
4	100	25	-32.91	-35.95	–	–	1.58	3.36
5	100	21	-36.60	-36.34	–	–	3.93	6.55

Table 2.A.2: Simulated estimation bias (percent underestimation) for non-truncated, non-sequential and all-study strategies in meta-analyses of $M = 24$ studies, each with 90% power for an effect size of $\delta/\sigma = 0.25$ at a one-sided significance level of 2.5%. All studies used the O'Brien-Fleming stopping boundary with K_m equally space interim analyses. Values in parentheses for the non-sequential strategy are the efficiencies (percent) relative to the all-study strategy. S and N are the percent of studies with sequential monitoring and non-truncation, respectively.

K_m	S	N	Estimation bias					
			Non-truncated strategy		Non-sequential strategy		All-study strategy	
			Fixed	Random	Fixed	Random	Fixed	Random
1	0	100	-0.19	-0.18	-0.19 (100)	-0.18 (100)	-0.19	-0.18
2	0	100	-0.40	-0.39	-0.40 (100)	-0.39 (100)	-0.40	-0.39
3	0	100	0.41	0.41	0.41 (100)	0.41 (100)	0.41	0.41
4	0	100	0.09	0.10	0.09 (100)	0.10 (100)	0.09	0.10
5	0	100	0.02	0.01	0.02 (100)	0.01 (100)	0.02	0.01
1	25	92	-1.70	-1.69	0.21 (81)	0.21 (82)	0.38	0.49
2	25	85	-2.55	-2.55	-0.03 (79)	-0.03 (79)	0.14	0.28
3	25	83	-2.34	-2.34	0.49 (83)	0.49 (84)	0.30	0.43
4	25	81	-2.01	-2.00	0.79 (76)	0.79 (77)	0.46	0.61
5	25	80	-2.51	-2.52	-0.15 (81)	-0.16 (81)	-0.05	0.10
1	50	84	-4.54	-4.54	-0.02 (55)	-0.02 (55)	0.15	0.44
2	50	70	-6.23	-6.22	0.29 (49)	0.31 (49)	0.79	1.17
3	50	65	-7.29	-7.30	-0.27 (57)	-0.28 (58)	-0.15	0.22
4	50	62	-6.50	-6.50	-0.03 (60)	-0.03 (60)	0.18	0.54
5	50	61	-6.50	-6.50	0.08 (56)	0.08 (57)	-0.36	-0.01
1	75	77	-7.59	-7.59	0.76 (27)	0.75 (28)	-0.13	0.41
2	75	55	-12.97	-12.96	-1.02 (35)	-1.01 (36)	0.57	1.25
3	75	48	-14.10	-14.11	-0.21 (30)	-0.21 (31)	-0.03	0.60
4	75	44	-14.22	-14.22	0.18 (26)	0.19 (26)	0.17	0.96
5	75	41	-13.44	-13.45	0.21 (30)	0.19 (31)	0.31	1.03
1	100	69	-10.93	-10.93	–	–	0.58	1.38
2	100	42	-22.73	-22.73	–	–	0.64	1.63
3	100	30	-30.34	-30.37	–	–	0.57	1.64
4	100	25	-34.98	-34.99	–	–	0.17	1.37
5	100	21	-37.62	-37.50	–	–	0.51	1.60

Table 2.A.3: Simulated estimation bias (percent underestimation) for non-truncated, non-sequential and all-study strategies in meta-analyses of $M = 12$ studies, each with 90% power for an effect size of $\delta/\sigma = 0.10$ at a one-sided significance level of 2.5%. All studies used the O'Brien-Fleming stopping boundary with K_m equally space interim analyses. Values in parentheses for the non-sequential strategy are the efficiencies (percent) relative to the all-study strategy. S and N are the percent of studies with sequential monitoring and non-truncation, respectively.

K_m	S	N	Estimation bias					
			Non-truncated strategy		Non-sequential strategy		All-study strategy	
			Fixed	Random	Fixed	Random	Fixed	Random
1	0	100	-0.13	-0.12	-0.13 (100)	-0.12 (100)	-0.13	-0.12
2	0	100	-0.93	-0.94	-0.93 (100)	-0.94 (100)	-0.93	-0.94
3	0	100	0.84	0.84	0.84 (100)	0.84 (100)	0.84	0.84
4	0	100	0.60	0.60	0.60 (100)	0.60 (100)	0.60	0.60
5	0	100	0.45	0.44	0.45 (100)	0.44 (100)	0.45	0.44
1	25	92	-0.77	-0.77	1.11 (80)	1.11 (81)	1.37	1.55
2	25	85	-2.87	-2.87	-0.22 (80)	-0.22 (80)	-0.05	0.13
3	25	83	-1.75	-1.75	1.16 (84)	1.16 (84)	0.66	0.85
4	25	81	-2.93	-2.93	-0.43 (83)	-0.43 (84)	-0.34	-0.07
5	25	81	-2.63	-2.63	-0.01 (80)	-0.01 (79)	-0.22	0.07
1	50	84	-4.39	-4.40	-0.24 (47)	-0.24 (48)	0.29	0.62
2	50	71	-6.51	-6.51	0.20 (49)	0.20 (50)	0.20	0.65
3	50	66	-6.09	-6.10	0.60 (62)	0.59 (65)	0.62	1.20
4	50	63	-6.74	-6.74	-0.46 (56)	-0.45 (56)	-0.23	0.22
5	50	61	-6.37	-6.38	0.06 (60)	0.06 (61)	0.08	0.66
1	75	76	-7.34	-7.34	1.06 (28)	1.06 (29)	0.42	1.12
2	75	56	-13.35	-13.35	-1.28 (26)	-1.27 (27)	0.13	1.07
3	75	48	-12.85	-12.85	1.28 (33)	1.02 (33)	0.68	1.60
4	75	45	-13.89	-13.90	0.35 (29)	0.35 (30)	0.00	0.97
5	75	42	-14.03	-14.03	-0.08 (31)	-0.07 (32)	0.13	1.13
1	100	71	-10.91	-10.91	–	–	0.14	1.26
2	100	42	-23.45	-23.50	–	–	0.25	1.43
3	100	31	-30.01	-29.62	–	–	0.02	1.27
4	100	25	-35.27	-35.31	–	–	0.01	1.50
5	100	22	-38.96	-38.99	–	–	0.00	1.58

Table 2.A.4: Simulated estimation bias (percent underestimation) for non-truncated, non-sequential and all-study strategies in meta-analyses of $M = 12$ studies, each with 80% power for an effect size of $\delta/\sigma = 0.25$ at a one-sided significance level of 2.5%. All studies used the O'Brien-Fleming stopping boundary with K_m equally space interim analyses. Values in parentheses for the non-sequential strategy are the efficiencies (percent) relative to the all-study strategy. S and N are the percent of studies with sequential monitoring and non-truncation, respectively.

K_m	S	N	Estimation bias					
			Non-truncated strategy		Non-sequential strategy		All-study strategy	
			Fixed	Random	Fixed	Random	Fixed	Random
1	0	100	-0.06	-0.07	-0.60 (100)	-0.07 (100)	-0.60	-0.07
2	0	100	-0.43	-0.42	-0.43 (100)	-0.42 (100)	-0.43	-0.42
3	0	100	-0.15	-0.16	-0.15 (100)	-0.16 (100)	-0.15	-0.16
4	0	100	0.76	0.76	0.76 (100)	0.76 (100)	0.76	0.76
5	0	100	0.43	0.42	0.43 (100)	0.42 (100)	0.43	0.42
1	25	94	-1.46	-1.47	0.43 (77)	0.42 (77)	0.62	0.79
2	25	89	-2.90	-2.91	0.29 (76)	0.29 (77)	0.13	0.38
3	25	86	-3.51	-3.49	-0.24 (77)	-0.23 (77)	-0.14	0.14
4	25	84	-1.81	-1.81	1.29 (82)	1.28 (83)	1.62	1.89
5	25	83	-2.30	-2.30	0.88 (80)	0.87 (80)	1.14	1.41
1	50	90	-3.21	-3.20	0.46 (61)	0.46 (62)	0.80	1.21
2	50	77	-6.94	-6.92	0.26 (53)	0.28 (55)	0.50	1.04
3	50	72	-8.54	-8.54	-1.02 (52)	-1.01 (54)	-0.17	0.44
4	50	68	-7.76	-7.77	-0.47 (55)	-0.47 (56)	0.81	1.31
5	50	67	-8.59	-8.57	-0.90 (54)	-0.87 (56)	0.17	0.86
1	75	84	-7.03	-7.03	1.79 (26)	1.81 (27)	-0.09	0.67
2	75	66	-11.99	-12.00	0.83 (32)	0.81 (32)	0.98	1.87
3	75	58	-15.30	-15.29	0.63 (29)	0.64 (31)	0.27	1.37
4	75	54	-15.64	-15.63	-0.17 (32)	-0.17 (34)	0.43	1.50
5	75	52	-14.99	-15.01	2.77 (35)	2.74 (37)	0.73	1.79
1	100	79	-9.57	-9.56	-	-	-0.14	1.02
2	100	56	-20.73	-20.72	-	-	0.52	1.90
3	100	44	-25.86	-25.84	-	-	1.29	2.57
4	100	38	-30.25	-30.56	-	-	1.56	3.16
5	100	35	-34.10	-34.51	-	-	0.48	2.22

2.A.5 Additional references

Durrett, R. (2010). *Probability: Theory and Examples*. Fourth. New York, New York: Cambridge University Press.

Muthén, B. (1990). Moments of the censored and truncated bivariate normal distribution. *British Journal of Mathematical and Statistical Psychology* **43**: 131–143.

Pechtl, A. (1998). A note on the derivative of the normal distribution's logarithm. *Archiv der Mathematik* **70**: 83–88.

Rosenbaum, S. (1961). Moments of a truncated bivariate normal distribution. *Journal of the Royal Statistical Society* **23**.B: 405–408.

Chapter 3

Treatment effect heterogeneity in subgroup analysis

Published as:

Schou, I. M. and I. C. Marschner (2015). Methods for exploring treatment effect heterogeneity in subgroup analysis: an application to global clinical trials. *Pharmaceutical Statistics* **14**: 44–55.

Abstract

Multi-country randomised clinical trials (MRCTs) are common in the medical literature and their interpretation has been the subject of extensive recent discussion. In many MRCTs, an evaluation of treatment effect homogeneity across countries or regions is conducted. Subgroup analysis principles require a significant test of interaction in order to claim heterogeneity of treatment effect across subgroups, such as countries in a MRCT. As clinical trials are typically underpowered for tests of interaction, overly optimistic expectations of treatment effect homogeneity can lead researchers, regulators and other stakeholders to over-interpret apparent differences between subgroups even when heterogeneity tests are insignificant. In this paper we consider some exploratory analysis tools to address this issue. We present three measures derived using the theory of order statistics which can be used to understand the magnitude and the nature of the variation in treatment effects that can arise merely as an artefact of chance. These measures are not intended to replace a formal test of interaction, but instead provide non-inferential visual aids allowing comparison of the observed and expected differences between regions or other subgroups, and are a useful

supplement to a formal test of interaction. We discuss how our methodology differs from recently published methods addressing the same issue. A case study of our approach is presented using data from the Study of Platelet Inhibition and Patient Outcomes (PLATO), which was a large cardiovascular MRCT that has been the subject of controversy in the literature. An R package is available that implements the proposed methods.

Keywords: clinical trial; heterogeneity; interaction; multi-country study; subgroup analysis

3.1 Introduction

Multi-country randomised clinical trials (MRCTs) evaluating new drug therapies are popular as they efficiently pool resources to provide faster recruitment and more generalisable results across patient populations, ethnicities and disease management paradigms. MRCTs also have the advantage of providing country-specific data that can be used for local regulatory dossiers that may otherwise require bridging studies, and local reimbursement applications for countries that have government funded pharmaceutical schemes. As a supplement to the overall analysis, MRCTs often present country-specific results that effectively correspond to a subgroup analysis. These subgroups may be defined by the individual countries participating in the study, or by pooling several countries in a geographical region, to avoid issues of low power or analytical complications that can arise from low enrolment in individual countries. In this paper we will focus on the interpretation of these country- or region-specific subgroup analyses, and will use the terms country and region interchangeably.

Clinical trials often assess the consistency of the treatment effect across pre-specified subgroups and generally accepted principles of subgroup analysis have been developed (Rothwell, 2005; Wang et al., 2007). In MRCTs, there is typically an assessment of treatment effect homogeneity across subgroups defined by regions. According to subgroup analysis principles, a test of interaction is the standard assessment of treatment effect heterogeneity across subgroups. However, as most studies are only designed with adequate power to

detect an overall clinically meaningful difference between treatments in the primary endpoint, the test of interaction to assess heterogeneity of treatment effect across regions in a MRCT can often be underpowered (ICH, 1998; Wittes, 2013). Indeed, the power decreases further as the number of regions in the subgroup analysis increases. Therefore, when there is a non-significant p-value from a test of interaction in the presence of seemingly heterogeneous treatment effects across regions, speculation of a type II error can arise making interpretation of the regional results difficult. It is very important that such speculation takes due account of the fact that random variation can result in some regions showing a lack of benefit even when there is no underlying heterogeneity and the treatment effect is beneficial. To this end, it is prudent to investigate whether potential differences exist between regions that can plausibly lead to differential treatment benefit and to appropriately design a study with this in mind (Hung, Wang, and O'Neill, 2010). A design paper or the study analysis plan can also be used to pre-emptively document and help inform researchers of the extent of chance variation to anticipate in a planned MRCT (Marschner, 2010).

Consideration of potential differences between region-specific treatment effects is important at both the design stage and the analysis stage of a MRCT. At the design stage it is useful to understand the nature and extent of chance differences that can be expected to arise between regions, under the assumption of treatment effect homogeneity. At the analysis stage it is useful to compare the observed regional treatment differences with the expected regional treatment differences and assess the magnitude of any differences. As such, the intent of this approach is not to determine how the methodology performs under heterogeneity. Instead, it assesses the potential extent of chance variation under an assumption of homogeneity. A previous paper focused on considerations at the design stage (Marschner, 2010), and in the present paper we adapt and extend this approach for application at the analysis stage. Our methods are based on the theory of order statistics for heteroscedastic normally distributed variables, which is applied to the collection of region-specific treatment differences. This allows various comparisons of expected subgroup-specific effects to be made with the actual observed effects under an assumption of treatment effect homogeneity. Specifically,

we investigate the expected and observed effects via a comparison of order statistics, the probability of subgroups favouring the control, and the distribution of the range of treatment effects. The resulting collection of graphical presentations provides a useful supplementary tool to the test of interaction and can equip researchers with a visual summary of the concordance between the observed treatment differences across regions and those expected due to chance. Although we will focus on regional differences in MRCTs, the methodology that we propose is equally applicable to other subgroup analyses.

Over recent years there has been a high level of research activity on statistical considerations relating to treatment effect heterogeneity in MRCTs and multi-centre studies, reflecting the practical importance of this issue (Wittes, 2013; Hung, Wang, and O'Neill, 2010; Marschner, 2010; Chen et al., 2010; Quan et al., 2010; Chen et al., 2011; Gallo et al., 2011; Ibia and Binkowitz, 2011; Chen et al., 2013). In the next section we will begin by reviewing past methods of relevance to those discussed here, including a very recently published approach which, like ours, is based on the theory of order statistics (Chen et al., 2013). We then introduce our methodological extensions, as well as providing a discussion of the fundamental differences between our approach and previous approaches, particularly our use of absolute treatment effects rather than standardised treatment effects in assessing the concordance between observed and expected treatment effect heterogeneity. Although this introduces methodological complexities compared to past approaches (Chen et al., 2013), we argue that this leads to more interpretable exploratory analysis tools. Finally we consider a detailed case study of the methods based on the PLATO study, which was a large MRCT of ticagrelor and clopidogrel for the prevention of cardiovascular events in patients with acute coronary syndromes (Wallentin et al., 2009). Application of our methods to the PLATO study, which has been the subject of much discussion in the literature, suggests that the apparently large variation in country-specific treatment effects is consistent with the play of chance.

3.2 Overview of previous research

We begin with an overview of previous work which our research extends, together with an introduction of the assumptions and notation that will be used throughout the paper.

3.2.1 Assumptions

Consider the comparison of two treatment groups, a control treatment group and an experimental treatment group, in a MRCT conducted over R regions. The sample size for treatment group i in region r is n_{ir}, for $i = 1,2$ and $r = 1,\ldots,R$. It is assumed that there is a parameter δ which measures the treatment effect, with $\delta = 0$ corresponding to no difference between the treatments. In principle the parameter δ could depend on r, meaning that there is genuine treatment effect heterogeneity across the regions. However, here we will make the assumption that δ does not depend on r, because our methods are aimed at assessing the extent of chance variation that could arise in the observed region-specific treatment effects under the assumption that there is underlying homogeneity.

The treatment effect δ could take a variety of forms depending on the type of primary endpoint that is being used in the MRCT. For example, with continuous endpoints δ may be a mean difference, with binary endpoints δ may be a risk difference, log relative risk or log odds ratio, while for time-to-event endpoints δ may be a log hazard ratio. Regardless of the type of treatment effect that δ measures, it will be assumed that for each region there is a region-specific estimator D_r of δ, which has a normal distribution

$$D_r \sim N\left(\delta, s_r^2\right) \qquad r = 1,\ldots,R. \tag{3.1}$$

This distributional assumption will be reasonable for most types of treatment effect measures on an appropriate scale, at least under a large sample assumption with approximate normality. Furthermore, it is assumed that the region-specific estimators are independent random variables. Other than these general assumptions it is not necessary for us to make any specific assumptions about the type of endpoint or the treatment effect measure δ. In the

case study described in Section 3.4 we will make use of the above model with a time-to-event endpoint where δ is a log hazard ratio parameter and D_r are country-specific log-hazard ratio estimators. However, it is also applicable for other treatment effect measures, and has been used for relative risks and risk differences in other contexts (Marschner, 2010; Schou and Marschner, 2013).

The form of s_r^2 in (3.1) can be derived in terms of the proportion of the study enrolment allocated to region r and the design parameters used in the overall sample size calculation, including the power, significance level and the homogeneous treatment difference δ. This form of s_r is useful for the assessment of expected treatment effect heterogeneity at the design stage, as illustrated by Marschner (2010). At the analysis stage s_r^2 will not be known in general, so a standard error estimate must also be available as discussed further in Section 3.3.5.

3.2.2 Expected range

Marschner (2010) proposed the expected range of region-specific treatment effects as a useful benchmark for the expected treatment effect variation. The expected range can be derived based on the distribution function of the smallest and largest order statistics, $D_{(1)}$ and $D_{(R)}$, which are respectively

$$F_{(1)}(x) = 1 - \prod_{i=1}^{R} \{1 - F_i(x)\} = 1 - \prod_{i=1}^{R} \left\{ 1 - \Phi\left(\frac{x - \delta}{s_i} \right) \right\}$$

and

$$F_{(R)}(x) = \prod_{i=1}^{R} F_i(x) = \prod_{i=1}^{R} \Phi\left(\frac{x - \delta}{s_i} \right).$$

Here, F_i is the distribution function of the normal distribution in equation (3.1) with $r = i$, while Φ is standard normal distribution function.

Using these distribution functions, the expectations of $D_{(1)}$ and $D_{(R)}$ can be calculated, as can the expectation of the range of treatment effects, $V = D_{(R)} - D_{(1)}$ (Marschner, 2010).

The expectation $E(V)$ provides a measure of the range of the treatment differences that can be expected due to chance, under an assumption of treatment effect homogeneity across the regions. The intent of this measure was to facilitate a comparison of the range of observed and expected treatment differences, thus providing a non-inferential complement to the primary assessment based on a test of interaction of treatment effect differences across regions. Subsequently the expected range has also been used in a more inferential capacity by Chen et al. (2013), although this was not the original intention.

3.2.3 Probability of at least one region favouring the control

An alternative measure that is also based on the extreme order statistics and provides information about the expected variation in region-specific treatment effects is the probability of at least one region favouring the control (Marschner, 2010; Li, Chuang-Stein, and Hoseyni, 2007). The motivation for considering this quantity is that an inconsistent region-specific treatment effect in a study that shows an overall benefit in favour of the experimental treatment will often prompt further investigation and interpretation. Quantifying the probability of this event, and the extent to which it is likely or unlikely, therefore provides a benchmark against which the occurrence of an inconsistent region-specific treatment effect can be interpreted.

Assuming δ is scaled such that a negative value for the treatment difference indicates benefit in favour of the experimental treatment, then the probability of at least one region favouring the control is given by

$$\Pr\left(D_{(R)} > 0\right) = 1 - \prod_{i=1}^{R} F_i(0) = 1 - \prod_{i=1}^{R} \Phi\left(\frac{-\delta}{s_i}\right).$$

As with the expected range, the intent of this measure is to provide a non-inferential tool to calibrate expectations about whether all treatment effects should lie in a consistent direction. If the probability is substantial, then it should not be too surprising if an inconsistent treatment effect is observed in a particular region, and over-interpretation of such an observation

should be avoided. Such information can be taken into consideration alongside the test of interaction.

3.2.4 Normal scores

While the extreme order statistics $D_{(1)}$ and $D_{(R)}$ provide information about treatment effect heterogeneity, it is natural to consider more informative methods based on all order statistics. A recently proposed alternative approach of Chen et al. (2013) does this. This approach assesses treatment effect heterogeneity using normal probability plots comparing the ordered standardised treatment differences with their associated normal scores. Specifically, the approach uses the standardised quantity referred to as the weighted least squares residual defined as $e_r = (D_r - \hat{\delta})/s_r$. Here

$$\hat{\delta} = \sum_{r=1}^{R} w_r D_r$$

is an unbiased estimator of δ with the weights $w_r = s_r^{-2} / \left(\sum_{i=1}^{R} s_i^{-2}\right)$ reflecting the amount of statistical information provided by region r, or equivalently the precision of the region-specific estimator D_r. Under the assumption of treatment effect homogeneity, the weighted least squares residuals are distributed as

$$e_r = \frac{(D_r - \hat{\delta})}{s_r} \sim N\big(0, 1 - w_r\big). \tag{3.2}$$

It then follows from (3.2) that the standardised weighted least squares residual $\widetilde{e}_r = e_r/\sqrt{1-w_r}$ has a standard normal distribution. The method proposes comparing the ordered standardised weighted least squares residuals $\widetilde{e}_{(r)}$, $r = 1, \ldots, R$, with the standard normal scores which can be readily obtained using standard tables or software (Arnold, Balakrishnan, and Nagaraja, 2008; R Development Core Team, 2014). The main tool for undertaking this comparison is a normal probability plot. In the special case of a homoscedastic normal outcome where δ is the mean difference and the treatment group sizes are equal within each region, the weights w_r reduce to the proportion of the overall

sample size that comes from region r (Chen et al., 2013). However, the above approach applies more generally, and can be used for other treatment difference measures that conform with the basic assumption (3.1).

In the present paper, our most significant contribution is to adapt this normal scores method to make use of the absolute order statistics $D_{(r)}$ in place of the standardised order statistics $\widetilde{e}_{(r)}$. In the next section we consider the substantial methodological complexities this introduces, but also explain why we believe this leads to a more interpretable assessment of treatment effect heterogeneity.

3.3 Methodological extensions

In this section we consider various extensions and adaptations of the methods reviewed in the previous section. We will focus on three measures that can be used in comparing the observed variation in treatment effects with what would be expected by chance under the assumption of treatment effect homogeneity across regions.

3.3.1 Overview of extensions

The first of the three measures we consider is the expected value of the r^{th} order statistic of the region-specific treatment effects, $E\big(D_{(r)}\big)$, for each $r = 1,\ldots,R$. Comparison of these expected order statistics with the sample order statistics $D_{(r)}$, for example using a normal probability plot, provides an alternative version of the comparison described in Section 3.2.4, between $\widetilde{e}_{(r)}$ and the normal scores. Although it may seem like a natural alternative to use of the absolute treatment effects rather than the standardised treatment effects, this introduces a number of complexities because the $D_{(r)}$ quantities are the order statistics from a heteroscedastic sample. These complexities are addressed in the next section. Despite the additional complexity we argue in Section 3.3.4 that this comparison provides a preferable assessment of treatment effect heterogeneity than the use of standardised treatment effects as used by Chen et al. (2013).

The second measure involves using the full distribution of the number of regions that favour the control, rather than the more restrictive quantity discussed in Section 3.2.3, the probability of at least one region favouring the control. This distribution will be helpful in interpreting studies where more than one region favours the control, which is not uncommon in MRCTs involving a large number of regions.

Finally, the third measure we consider is the full probability distribution of the treatment effect range, $D_{(R)} - D_{(1)}$, which is helpful in interpreting the treatment effect range observed in a MRCT. Use of the full distribution generalises the expected range approach described in Section 3.2.2, which is based just on the expected value of the effect range distribution. In principle this approach could also be generalised to other range-based distributions, such as the distribution of the inter-quartile range of region-specific treatment effects. Here, however, we restrict our focus to the range of treatment effects which, as described in Section 3.2.2, has been the focus of prior research.

3.3.2 Order statistic distribution

All of our methods depend fundamentally on the distribution of the order statistics of the region-specific treatment effects. This involves considering the order statistics from a sample of R heteroscedastic normal variates. We now consider this distribution and then describe how it can be used to derive the three measures of expected treatment effect heterogeneity.

The distribution function of $D_{(r)}$ is

$$\begin{aligned} F_{(r)}(x) &= \Pr\left(D_{(r)} \leq x\right) \\ &= \Pr\left(\text{At least } r \text{ of } R \text{ treatment differences do not exceed } x\right) \\ &= \sum_{i=r}^{R} \sum_{S \in S_i(R)} \left\{ \prod_{k \in S} F_k(x) \prod_{\substack{k=1 \\ k \notin S}}^{R} \left[1 - F_k(x)\right] \right\} \end{aligned} \tag{3.3}$$

where $S_i(R)$ is the family of all subsets of size i from $\{1, \ldots, R\}$ (Balakrishnan, 2007).

On expansion and simplification of (3.3) we get

$$F_{(r)}(x) = \sum_{i=r}^{R} c_{ir} \sum_{S \in S_i(R)} \prod_{k \in S} F_k(x) \tag{3.4}$$

where

$$c_{ir} = (-1)^{i-r} \binom{i-1}{r-1}.$$

In the special case where the D_rs are independent identically distributed random variables with $s_r = s$, equation (3.4) reduces to

$$F_{(r)}(x) = \sum_{i=r}^{R} c_{ir} \binom{R}{i} F(x)^i,$$

and is equivalent to the familiar representation (Arnold, Balakrishnan, and Nagaraja, 2008)

$$F_{(r)}(x) = \sum_{i=r}^{R} \binom{R}{i} F(x)^i \big[1 - F(x)\big]^{R-i}.$$

However, our formulation allows for a fully heteroscedastic specification which is required to allow for different regions having different sample sizes.

Applying the product rule for differentiation on the distribution function, the probability density of the r^{th} order statistic is

$$f_{(r)}(x) = \sum_{i=r}^{R} \sum_{j=1}^{R} c_{ir} f_j(x) \sum_{S \in S_i(R)} 1\{j \in S_i(R)\} \prod_{\substack{k \in S \\ k \neq j}} F_k(x) \tag{3.5}$$

where $1\{\cdot\}$ is the indicator function. Although this theoretical specification appears unwieldy, it is straightforward to compute.

As with $F_{(r)}(x)$, a simplified version of (3.5) is achieved in the special case where the D_rs are independent identically distributed random variables with $s_r = s$, and is given by

$$f_{(r)}(x) = \sum_{i=r}^{R} i c_{ir} \binom{R}{i} f(x) F(x)^{i-1}.$$

A simplified illustrative example of the order statistic distribution is provided for a MRCT with $R = 3$ regions and treatment differences D_1, D_2, and D_3. In this case, consider the distribution of $D_{(2)}$. Here, the family of sets $S_2(3)$ and $S_3(3)$ would be given by $S_2(3) = \{\{1,2\},\{1,3\},\{2,3\}\}$ and $S_3(3) = \{\{1,2,3\}\}$. The distribution and density functions of $D_{(2)}$ follow readily from (3.4) and (3.5) and the fact that $c_{22} = 1$ and $c_{32} = -2$, namely

$$\begin{aligned} F_{(2)}(x) &= F_1(x)F_2(x)\left[1-F_3(x)\right] + F_1(x)F_3(x)\left[1-F_2(x)\right] \\ &+ F_2(x)F_3(x)\left[1-F_1(x)\right] + F_1(x)F_2(x)F_3(x) \\ &= c_{22} \times \{F_1(x)F_2(x) + F_1(x)F_3(x) + F_2(x)F_3(x)\} \\ &+ c_{32} \times \{F_1(x)F_2(x)F_3(x)\} \end{aligned}$$

and

$$\begin{aligned} f_{(2)}(x) &= c_{22} \times \{f_1(x)F_2(x) + f_1(x)F_3(x) + f_2(x)F_1(x) \\ &\quad + f_2(x)F_3(x) + f_3(x)F_1(x) + f_3(x)F_2(x)\} \\ &+ c_{32} \times \{f_1(x)F_2(x)F_3(x) + f_2(x)F_1(x)F_3(x) + f_3(x)F_1(x)F_2(x)\}. \end{aligned}$$

The forms of $F_{(2)}(x)$ and $f_{(2)}(x)$ in this simplified 3-region example illustrate the link between equations (3.3) and (3.4) and the role of the c_{ir} constants in specifying the order statistic distribution.

As foreshadowed in Section 3.3.1, the general order statistic distribution for heteroscedastic treatment effects can now be used to derive several useful measures of chance treatment effect variation that extend and improve upon the measures discussed in Section 3.2.

3.3.3 Measures of chance variation

The first measure described in Section 3.3.1, the expectation of the r^{th} order statistic of the region-specific treatment differences, can now be obtained using (3.5)

$$E\left(D_{(r)}\right) = \int_{-\infty}^{\infty} x f_{(r)}(x) dx. \tag{3.6}$$

Although this form is not explicit, it can be straightforwardly computed using standard routines for numerical integration. As explained later in the paper, all computations presented here were performed in R (R Development Core Team, 2014).

Once computed, these expected order statistics can be compared graphically with the observed ordered treatment differences to assess whether the observed spread of region-specific treatment differences is unusual relative to what would be expected by chance under the assumption of treatment effect homogeneity. One such plot would be a simple box plot of the observed and expected order statistics which provides a graphical generalisation of the approach of comparing the observed and expected ranges (Marschner, 2010). Another approach would be a plot of the observed versus expected treatment differences, which is a type of normal probability plot that conveys information about any departures of the observed region-specific treatment effects from what would be expected by chance.

Treatment differences that align consistently across regions in terms of direction of effect are straightforward to interpret and explain. Sometimes, however, chance variation will lead some regions to have a treatment effect estimate that goes in the opposite direction to the overall effect. This can potentially lead to speculation and over-interpretation. Therefore, being able to compare the observed number of regions favouring the control with the probability distribution of the number of regions favouring the control provides a useful benchmark by which to assess the role of chance variation. This leads to the second of the three approaches introduced in Section 3.3.1, which generalises the previously suggested approach discussed in Section 3.2.3.

Like the other quantities discussed in this section, the probability distribution of W, the number of regions favouring the control, is connected to the order statistic distribution discussed in Section 3.3.2 through the relationship

$$\Pr(W \geq w) = \Pr\big(D_{(R-w+1)} > 0\big) = 1 - F_{(R-w+1)}(0) \qquad w = 1, \ldots, R. \tag{3.7}$$

Assuming a positive treatment difference signifies an effect in favour of the control treatment, and letting p_i be the probability that region i favours the control, we have the following

$$p_i = \Pr\big(D_i > 0\big) = 1 - F_i(0) = 1 - \Phi\left(-\delta/s_i\right).$$

It then follows from equations (3.3) and (3.7) that the probability function of the number of regions favouring the control is

$$\begin{aligned} P_W(w) &= \Pr(W = w) = \Pr(W \geq w) - \Pr(W \geq w+1) \\ &= F_{(R-w)}(0) - F_{(R-w+1)}(0) \\ &= \sum_{S \in S_w(R)} \prod_{k \in S} p_k \prod_{\substack{l=1 \\ l \notin S}}^{R} (1 - p_l) \qquad w = 0, \ldots, R \end{aligned} \tag{3.8}$$

where, for notational purposes, we define $F_{(0)}(0) = 1$ and $F_{(R+1)}(0) = 0$. For example, in the 3-region illustration discussed previously, the probability that two regions favour the control is

$$P_W(2) = \Pr(W = 2) = p_1 p_2 (1 - p_3) + p_1 p_3 (1 - p_2) + p_2 p_3 (1 - p_1).$$

Once this distribution has been computed, the observed number of regions favouring the control can be compared with $P_W(w)$ in order to assess whether the observed number is unusual compared with what would be expected by chance under the assumption of homogeneous treatment effects. A natural summary measure of the extent to which the observation $W = w_o$ is consistent with chance variation, is the probability of obtaining an observation at least as extreme as $W = w_o$, namely, $P_E = \Pr\big(W \geq w_o\big)$. Although we are not recommending P_E as a p-value for formal hypothesis testing, it does nonetheless provide a non-inferential

quantification of the extent to which the observed number of inconsistent regions is unusual relative to what would be expected by chance.

Finally, the third approach introduced in Section 3.3.1 is based on the probability distribution of the range of region-specific treatment effects, $V = D_{(R)} - D_{(1)}$. This distribution is well known in the homoscedastic case based on the joint distribution of $D_{(1)}$ and $D_{(R)}$ (Arnold, Balakrishnan, and Nagaraja, 2008). In the heteroscedastic generalisation that we are using in this paper, the density function of the range can be expressed as follows for $x \geq 0$.

$$f_V(x) = \int_{-\infty}^{\infty} \sum_{i=1}^{R} \sum_{\substack{j=1 \\ j \neq i}}^{R} f_i(y) f_j(y+x) \prod_{\substack{k=1 \\ k \neq i,j}}^{R} [F_k(y+x) - F_k(y)]\, dy. \tag{3.9}$$

As in equation (3.6), equation (3.9) requires numerical integration which we have undertaken in R using the `integrate` routine (R Development Core Team, 2014). Once computed, the observed range can be compared with the probability distribution $f_V(x)$ to assess whether the observed range of treatment effects is unusual relative to what would be expected by chance under an assumption of treatment effect homogeneity. As with W above, a natural summary measure of the extent to which the observation $V = v_o$ is consistent with chance variation, is provided by the probability of obtaining an observation at least as extreme as $V = v_o$, which in this case is $P_E = \int_{v_o}^{\infty} f_V(x)dx$.

An R package called `subgroup`, that implements all three approaches, is available for download from the Comprehensive R Archive Network (CRAN) (Schou, 2014).

3.3.4 Comparison of the methods

While our methods and those of Chen et al. (2013) both make use of assessments that are based on the theory of order statistics, there are important differences between the two approaches. Most significantly, our approach uses the observed region-specific treatment differences whereas the approach proposed by Chen et al. (2013) uses the standardised treatment differences in the form of the weighted least squares residuals. In view of these differences, a discussion of the distinction between the two approaches is necessitated.

Statistically, the key distinction between using the absolute order statistics $D_{(r)}$ and the standardised order statistics $\widetilde{e}_{(r)}$, is that the former depends only on the treatment effects themselves, while the latter depends on a combination of the departure of the treatment effects from the overall effect and the associated standard error. Therefore, an ordering of the standardised weighted least squares residuals is essentially an ordering of the departure of the treatment effects from the overall effect, relative to the region-specific standard error, with the size of the standard error playing a critical role in the ordering. This may mean that the $D_{(r)}$ and $\widetilde{e}_{(r)}$ values are ordered in different ways. Indeed, this may mean that the two versions of order statistics convey different messages about whether the observed region-specific treatment effects are consistent with what would be expected by chance, and we provide an example of this in the case study discussed in Section 3.4.

The fact that the absolute and standardised treatment effects can convey different messages makes it important to consider how subgroup analyses are interpreted and used in practice by stakeholders. While the standardised treatment effects are what drives the formal test of heterogeneity, they are not the primary focus of subsequent informal assessments of the region-specific differences in treatment effects. Such informal assessments, which would typically follow an insignificant test of heterogeneity, tend to focus on the absolute magnitudes of the treatment difference in each region. The spread in these absolute treatment effects is what then has the potential to lead to over-interpretation of apparent treatment effect variation. It therefore makes sense to focus on the expected variation in absolute treatment effects as a benchmark for the observed variation in absolute treatment effects. It is this use of actual rather than standardised treatment effects in the assessment and interpretation of heterogeneity that has led us to base our measures of expected variation on the actual treatment effects.

3.3.5 Implementation issues

In practice, there are several implementation issues that we discuss prior to considering a case study. Firstly, it requires noting that the various quantities discussed in the previous

section are dependent on the unknown values of δ and s_r. This means that at the analysis stage of a study, sample estimates $\hat{\delta}$ and $\hat{s}_i$ are required so that computations of the expected variation in treatment effects can be undertaken. If the individual patient data are available, the overall treatment effect estimated using these data would be the most appropriate estimate of δ, as an assumption of treatment effect homogeneity underpins the assessment of chance variation. However, if only region-specific treatment effect estimates are available, the aggregated estimate of δ, as discussed in Section 3.2.4, would be used.

With regards to the standard error s_i, there are two possible approaches to estimation. The first, as used in this paper, would be to use the standard errors of the region-specific treatment effects as estimated separately within each region. This provides a more empirical estimate of standard error than the second approach which is to use the overall estimate of standard error, weighted by the proportion of subjects from each region. The latter approach enforces an assumption of region-level homoscedasticity and results in smaller regions being weighted less and larger regions being weighted more. This is a more natural approach to take at the design stage when no data is available.

A further implementation issue relates to the computational complexity of the methods. In Section 3.3.3 we presented theoretical expressions associated with the various measures of heterogeneity, that can be computed exactly with the aid of a routine to undertake numerical integration. In practice, it is also possible to approximate all of the required quantities using simulation. Although this is potentially computationally expensive, the computations themselves are trivial and obvious with the availability of a large number of simulated samples $D_1,\ldots,D_R$ from the normal distributions $N\left(\hat{\delta},\hat{s}_r^2\right)$, for $r = 1,\ldots,R$. Since the theoretical computations required to compute the quantities described in Section 3.3.3 are based on combinatorial sets, there will generally be a point at which simulation becomes more efficient than direct computation. Based on our experience with the case study described in Section 3.4, the simulation approach tends to be preferable for $R > 20$.

3.4 Case study

3.4.1 PLATO study

As a case study, we consider the Study of Platelet Inhibition and Patient Outcomes (PLATO), which was a 43-country, double-blind, randomised trial comparing the experimental treatment ticagrelor with the control treatment clopidogrel, for the prevention of cardiovascular events in 18,624 subjects with acute coronary syndrome (Wallentin et al., 2009). The primary endpoint of this study was the time to first occurrence of a cardiovascular event (death from vascular causes, myocardial infarction, or stroke). The study was designed to have 90% power to detect a relative risk reduction of 13.5%.

On completion, the overall study showed a significant reduction in cardiovascular events in favour of ticagrelor (hazard ratio 0.84, $p < 0.001$). Treatment effect heterogeneity was assessed in 33 separate subgroup analyses, one of which was an assessment of the heterogeneity of treatment effects across regions (Asia/Australia, Central/South America, Europe/Middle East/Africa and North America). The p-value for this test of interaction was 0.045 with the treatment effect in North America having an observed value that favoured the control, although insignificantly so. The investigators concluded that this finding may have been a chance result due to multiple testing, and that although no apparent explanations had been found, questioned whether the differences between patient populations and treatment practice patterns may have contributed to this result.

Although a p-value of 0.045 in the context of 33 subgroup analyses is not particularly surprising, the PLATO study was subsequently subjected to extensive post hoc analysis of country-specific heterogeneity in treatment effects. These analyses focused particularly on the observation that the USA treatment effect was in the direction favouring the control. The Food and Drug Administration (FDA) conducted its own review of the data following the sponsor's proposal of a potentially negative association between the dose of aspirin and the benefit of treatment with ticagrelor, finding that the dose of aspirin was higher in the USA subgroup compared with the non-USA subgroup (FDA, 2010). A further review

of this possible explanation was subsequently published together with a claim that differences in primary site monitoring by an independent contract research organisation (in the USA) and the study sponsor (in most other countries) may offer an alternative explanation requiring further investigation (Serebruany, 2010). These proposals of a potential biological explanation (aspirin dose) and an operational explanation (site monitoring) were followed by a statistical assessment concluding that the country-specific treatment effect variation was consistent with the play of chance (Buyse and Marschner, 2011) and a further analysis concluding that the findings in the USA were likely not due to chance (Chen et al., 2013). Here we use our methods to provide further exploration of the play of chance as a potential explanation for country-specific treatment effect differences in the PLATO study.

3.4.2 Data and analyses

In our analyses, we used published country-specific hazard ratios and 95% confidence intervals for all countries except the smallest (Hong Kong), which had only 16 patients. This led to $R = 42$ countries with sample sizes varying from 51 to 2666. We refer the reader to Figure 1 of Serebruany (2010) for a full listing of the countries, sample sizes and hazard ratios used in our analyses. The overall treatment effect δ was taken to be the log hazard ratio, for which assumption (3.1) is reasonable. The overall estimate $\hat{\delta}$ was calculated using an inverse variance weighting method based on country-specific log hazard ratios and standard errors calculated from the published confidence intervals.

As well as analyses of the treatment effects for all 42 countries, we also considered analyses restricted just to the countries with the largest sample sizes. These additional analyses served two purposes. Firstly, they enabled an assessment of the extent to which any conclusions are robust to the larger variation expected in small countries, which was raised as a concern by Chen et al. (2013). Secondly, these analyses served to illustrate the behaviour of the methodology on data sets having various R values. In our analyses we consider the results restricted to the largest 10, 15 and 20 countries, in addition to the full collection of 42 countries.

3.4.3 Order statistics

Figures 3.1 and 3.2 present the expected order statistics of the country-specific treatment differences displayed as box plots and normal probability plots. These plots are displayed for the entire collection of 42 countries, as well as analyses restricted to the largest 10, 15 or 20 countries. Also shown, in Figure 3.2 Panels B and D, are normal probability plots corresponding to the standardised weighted least squares residuals of Chen et al. (2013), as discussed in Section 3.2.4. Since formal tests of heterogeneity of treatment effects are statistically insignificant ($p > 0.1$ in all cases), we intend that these graphical displays are used as a non-inferential supplement to a formal test of heterogeneity, in which the observed variation in treatment effects is compared with the expected variation in treatment effects. With this in mind, these figures do not identify any remarkable differences between what was observed and what would be expected due to chance variation. Figure 3.1 clearly displays the expected increase in treatment effect variation as more countries are included in the analysis, but does not suggest that the observed variation is inconsistent with what was expected under the hypothesis of homogeneity. Indeed, for the analyses involving larger numbers of countries (Panels C and D) it appears that the PLATO study actually exhibits less variation in country-specific treatment effects than would have been expected due to chance. This is also evident in Figure 3.2 Panel C, where the shallow gradient for all but the most extreme order statistics is indicative of smaller variation than expected.

Of particular interest is the comparison of Panels A and B of Figure 3.2, which is a comparison of the normal probability plots for absolute treatment effects and standardised treatment effects, for the analysis restricted to the largest 15 countries. Panel A, based on absolute treatment effects, displays no departure from the expected variation of treatment effects, with the possible exception of the smallest order statistics that suggest lower variation than expected. On the other hand, the standardised treatment effects displayed in Panel B show one outlying country, the USA, which seems to have a standardised treatment effect that departs from the other countries. This illustrates the potential for different qualitative messages to emerge from these two methods.

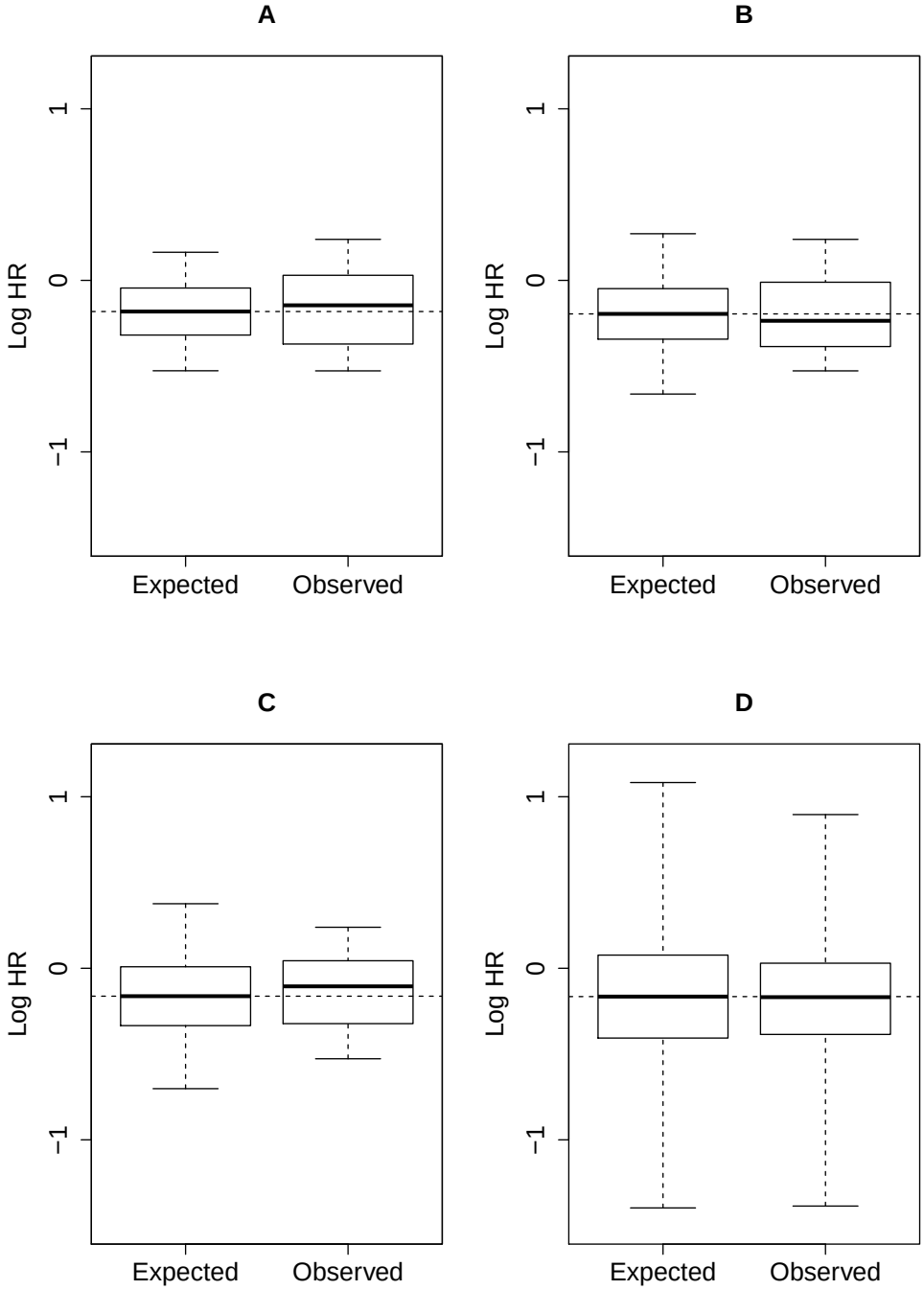

Figure 3.1: Observed and expected country-specific treatment differences from the PLATO study. The expected treatment differences for the largest 10 (Panel A), 15 (Panel B), 20 (Panel C) and 42 (Panel D) countries are plotted. The dotted line denotes the overall observed treatment difference.

3.4.4 Range of treatment effects

The expected range of treatment effects depicted in the extremities of the boxplots in Figure 3.1 can be generalised to the full distribution of the range of treatment effects, as discussed

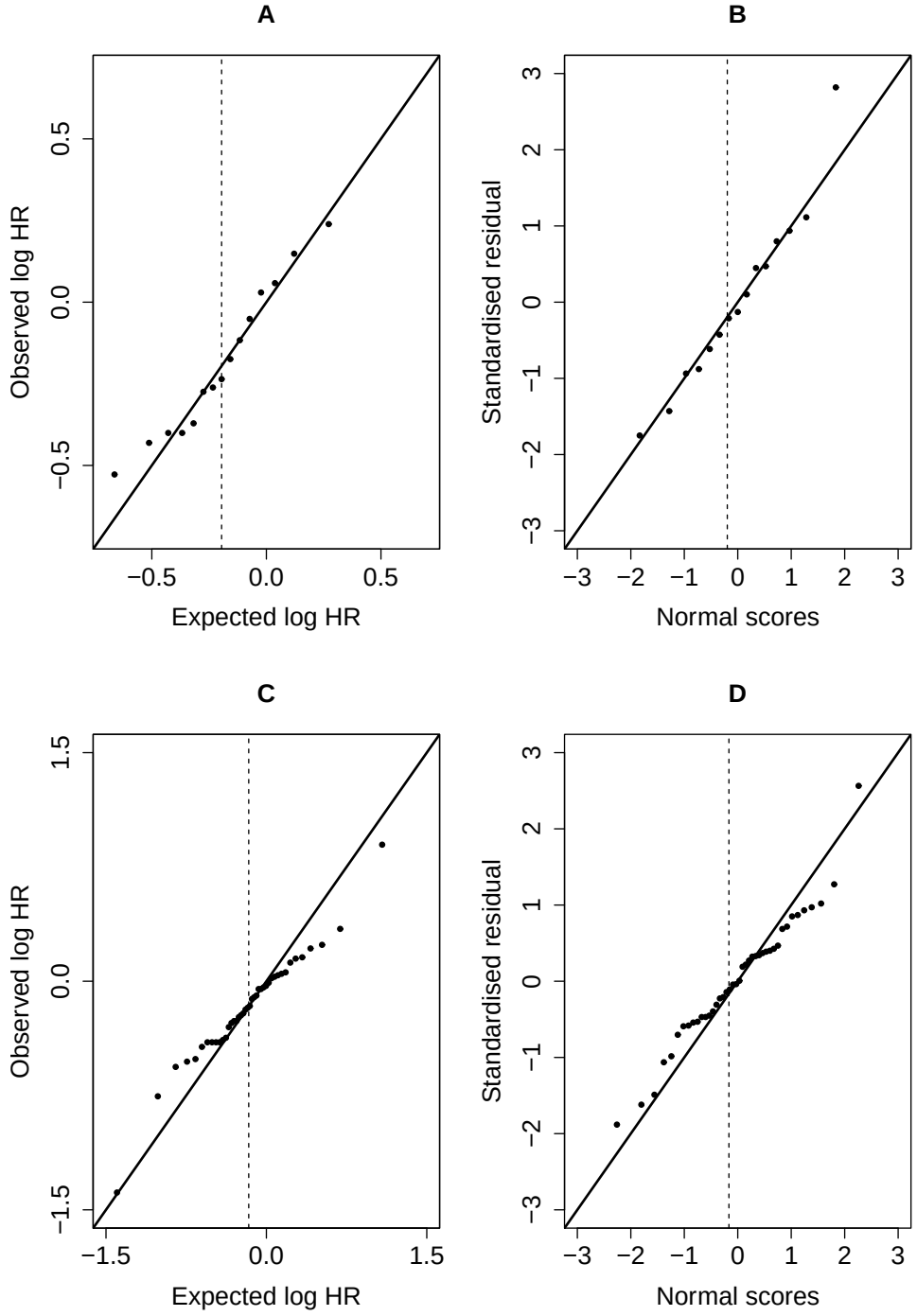

Figure 3.2: Observed and expected treatment differences from the largest 15 (Panel A and B) and 42 (Panel C and D) countries in the PLATO study. Panels A and C use absolute treatment effects whereas Panel B and D use the standardised weighted least squares residuals.

in Section 3.3.3. Plots of this distribution, together with the observed range of treatment effects, are provided in Figure 3.3. It can be seen that the observed range of country-specific treatment effects in the PLATO study is highly consistent with the distribution of the range

of treatment effects under the assumption of treatment effect homogeneity. This conclusion is true regardless of whether analyses are restricted to the largest countries or include all countries. A useful summary measure of the extent of consistency is P_E, which was described in Section 3.3.3. In the present context, P_E is the probability of observing a treatment effect range at least as extreme as the one observed, under the assumption of treatment effect homogeneity. With $P_E = 0.55$, the overall analysis in Panel D of Figure 3.3 shows that a treatment effect range as large as the one observed in the PLATO study is highly likely, and could therefore plausibly have arisen through chance variation. The same conclusion would also be reached using the P_E values restricted to the largest countries, as displayed in Panels A–C of Figure 3.3.

As a supplement to Figure 3.3, in Figure 3.4 we have displayed the observed and expected range of country-specific treatment effects for analyses restricted to the largest R countries, where R ranges from 10 through 42. It is clear from Figure 3.4 that regardless of whether the expected range of treatment effects is restricted to just the very large countries, or whether it includes the smaller countries with larger expected variation, the observed range of treatment effects is always consistent with the expected range.

3.4.5 Countries favouring the control

One feature that often causes concern in MRCTs with an overall experimental treatment benefit, is the occurrence of inconsistent country-specific treatment effects; that is, one or more country-specific treatment effects in the direction favouring the control treatment. This was certainly a concern in PLATO, particularly because one of these countries was the USA. In a study with as many countries as PLATO and a moderate overall treatment benefit, it is virtually certain that at least one country will have a treatment effect favouring the control, even if the treatment effect is homogeneous across countries. However, PLATO had 12 countries out 42 with treatment effects favouring the control, and when restricted to the largest countries, had 7 inconsistent effects out the largest 20 countries, 4 inconsistent effects out of the largest 15 countries, and 3 inconsistent effects out the largest 10 countries.

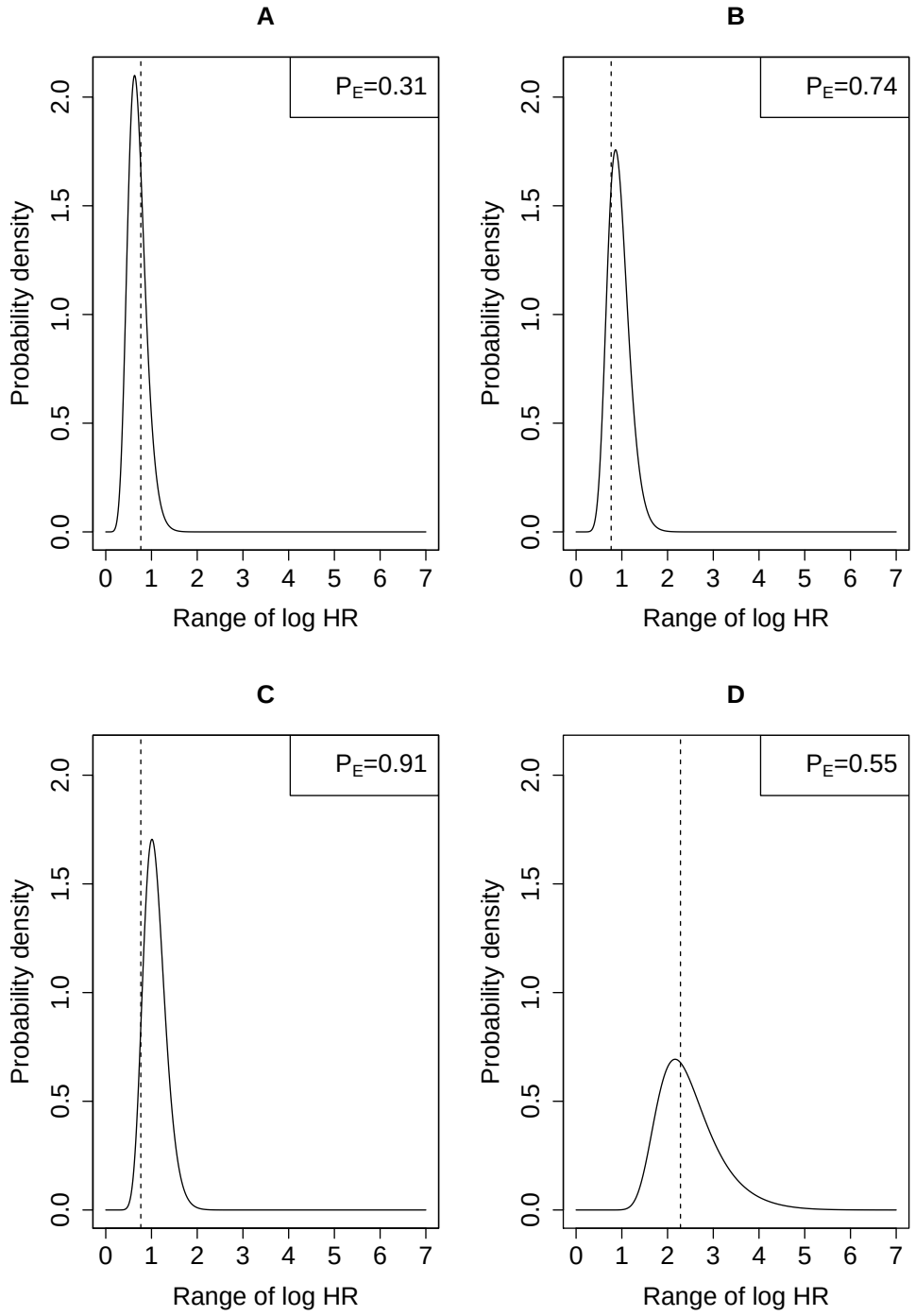

Figure 3.3: Probability density of the treatment effect range for the largest 10 (Panel A), 15 (Panel B), 20 (Panel C) and 42 (Panel D) countries in the PLATO study. The dotted line denotes the observed range.

These numbers of inconsistent countries may seem large, but when benchmarked against the probability distribution of the number of treatment effects favouring the control, as described in Section 3.3.3, it can be seen that they are not unusually large. Figure 3.5 displays

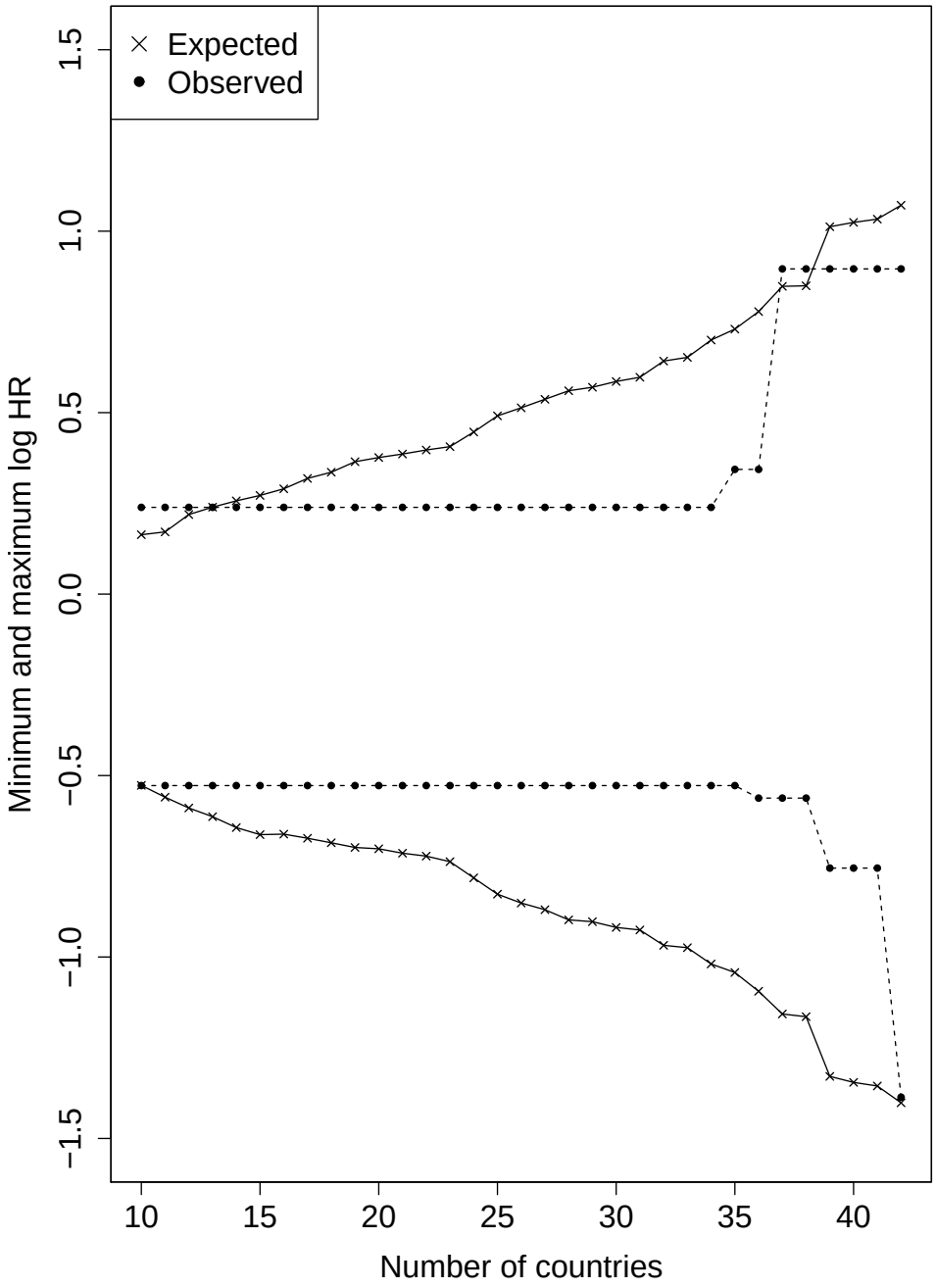

Figure 3.4: The range of observed and expected country-specific treatment effects in the PLATO study, restricting the analysis to the largest R countries, where R ranges from 10 to 42.

these distributions, together with the observed numbers of inconsistent countries, and the summary measure P_E which is the probability of an observation at least as extreme as the one observed. With $P_E = 0.72$ for the overall analysis in Panel D of Figure 3.5, it can be

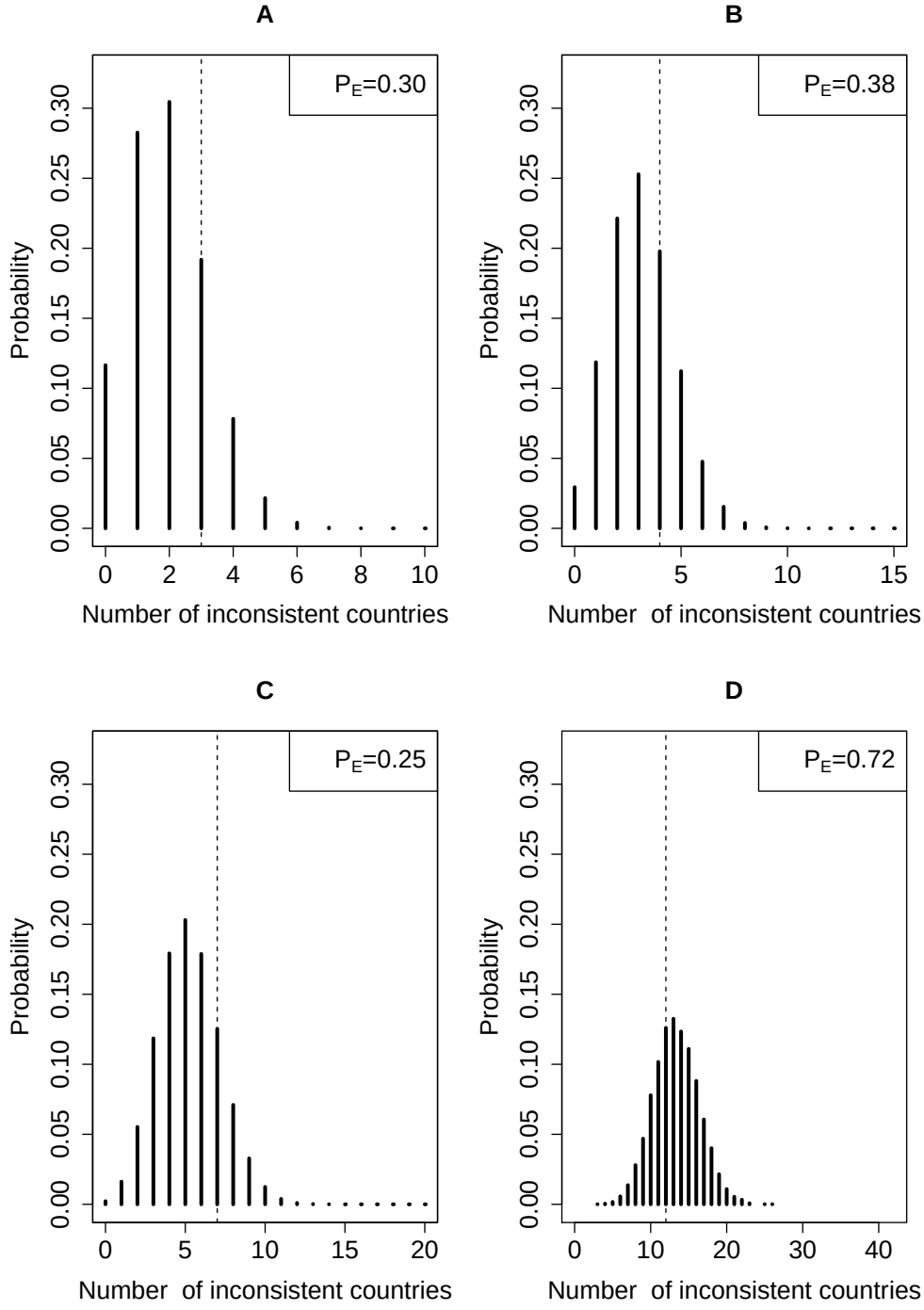

Figure 3.5: Probability distribution for the number of countries favouring the control for the largest 10 (Panel A), 15 (Panel B), 20 (Panel C) and 42 (Panel D) countries in the PLATO study. The dotted line denotes the observed value.

seen that an observation of 12 or more inconsistent countries is highly likely even under the assumption of treatment effect homogeneity. This conclusion is not altered by restricting the analysis to the largest countries, as in Panels A–C of Figure 3.5, all of which also show

that the observed number of inconsistent countries is not unusual relative to what would be expected by chance. Thus, these analyses suggest that any speculation about the causes of inconsistent country-specific treatment effects in PLATO, should acknowledge chance variation as a highly plausible explanation.

3.4.6 Conclusions

Despite all of the post hoc analysis and interpretation that the PLATO study has been subjected to, we conclude from our results that there is nothing particularly remarkable about the spread of treatment effects across countries. In a global study as large as the PLATO study, with over 40 countries, it is to be expected that wide variation in treatment effects will be observed. Consistent with earlier more limited analyses (Buyse and Marschner, 2011), our methods provide a suite of presentations suggesting that chance variation is a very plausible explanation for the spread of country-specific treatment effects observed in the PLATO study.

Finally, we note that our analyses were repeated to investigate how the various measures changed when a proportionally weighted overall standard error was used to estimate the s_r standard errors, as discussed in Section 3.3.5, instead of the individual country-specific standard errors used in the above analyses. It was found that there was very little difference between this approach and the approach presented in this section for the PLATO study.

3.5 Discussion

Assessment of heterogeneity of treatment effects between subgroups is a key element of clinical trial analysis. Recently, subgroup analysis of regional differences in MRCTs has become a prominent issue in the literature. In this paper we provide some new tools that aid interpretation of subgroup-specific treatment effects, and have illustrated these using a case study from a MRCT.

When a test of interaction is underpowered, and treatment effects are seemingly different

between subgroups, speculation may arise that there is heterogeneity of treatment effects that was not detected by the test of interaction. The approach we propose here is a non-inferential supplement to a formal test of interaction. A non-inferential approach has been suggested given that the same limitation of low power for a test of interaction will likely affect any new inferential technique one might develop to assess treatment effect heterogeneity. The suite of graphical tools introduced in this paper provide a multi-faceted visual assessment of the extent to which the observed treatment differences align with those that would be expected under an assumption of treatment effect homogeneity. That is, the intent is not to assess how these methods will perform under heterogeneity, but rather to quantify the potential extent of variation resulting from the play of chance under an assumption of homogeneity. Given the attention heterogeneity of treatment effects across regions has received in some MRCTs (Wallentin et al., 2009; Serebruany, 2010; Wedel et al., 2001), our approach provides additional tools for evaluating the extent of chance variation expected in a MRCT, and can be used to benchmark expectations and pre-empt any over-interpretation. The graphical nature of our methods make it amenable for interpretation by all stakeholders including non-statisticians.

The proposed methods supplement a formal test of heterogeneity by quantifying the extent of chance variation that is consistent with homogeneity. They are not intended as new methods to actually detect heterogeneity, and they do not provide a way to overcome a low powered test of heterogeneity. In addition to statistical analysis, assessment of the plausibility of heterogeneous treatment effects requires critical examination of the study design, data collection methods, treatment administration methods, biological mechanisms and other factors.

Treatment differences in typical clinical trial subgroups such as age and sex may present a plausible biological mechanism that explains the difference. However, treatment differences between regions are often more complex to understand because region is a composite of many variables that can potentially influence the outcomes of an intervention (Wittes, 2013). Thorough evaluation of potential treatment differences between regions at the design stage

of a study is critical, and can assist with the interpretation of any apparent heterogeneity that emerges at the analysis stage.

Our methods differ from a recently published method by Chen et al. (2013) in that we use the observed treatment differences whereas Chen et al. (2013) use the standardised treatment differences as defined by the weighted least squares residuals. Although this difference may seem trivial, the results and their interpretation can be quite different as the ordering proposed by Chen et al. (2013) depends on the relative magnitude of the departure of the region-specific treatment effect from the overall effect, compared with its standard error. We advocate the use of the observed treatment differences as these are required in practice for such activities as cost-effectiveness analyses and risk stratification in addition to the direct relevance they have for the physician and the patient.

In conclusion, our methods provide a non-inferential yet visually informative summary of the subgroup-specific variation in treatment effects that can arise as an artefact of chance. The appeal of these methods is their broad applicability, not just to global clinical trials as discussed here but also to other types of subgroup analysis, as well as the accessibility of the visual displays to all stakeholders including non-statisticians.

3.6 References

Arnold, B. C., N. Balakrishnan, and H. N. Nagaraja (2008). *A First Course in Order Statistics*. Philadelphia, USA: Society for Industrial and Applied Mathematics.

Balakrishnan, N. (2007). Permanents, order statistics, outliers, and robustness. *Revista Matemática Complutense* **1**: 7–107.

Buyse, M. and I. C. Marschner (2011). Assessment of statistical heterogeneity in the PLATO trial. *Cardiology* **118**: 138.

Chen, J., H. Quan, B. Binkowitz, S. P. Ouyang, Y. Tanaka, G. Li, S. Menjoge, and E. Ibia for the Consistency Workstream of the PhRMA MRCT Key Issue Team (2010). Assessing consistent treatment effect in a multi-regional clinical trial: a systematic review. *Pharmaceutical Statistics* **9**: 242–253.

Chen, J., H. Quan, P. Gallo, S. Menjoge, X. Luo, Y. Tanaka, G. Li, S. P. Ouyang, B. Binkowitz, E. Ibia, S. Talerico, and K. Ikeda (2011). Consistency of treatment effect across regions in multiregional clinical trials, part 1: design considerations. *Drug Information Journal* **45**: 595–602.

Chen, J., H. Zheng, H. Quan, G. Li, P. Gallo, S. P. Ouyang, B. Binkowitz, N. Ting, Y. Tanaka, X. Luo, and E. Ibia for the Society for Clinical Trials (SCT) Multi-Regional Clinical Trial Consistency Working Group (2013). Graphical assessment of consistency in treatment effect among countries in multi-regional clinical trials. *Clinical Trials* **10**: 842–851.

FDA (2010). Ticagrelor Secondary Review. http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/CardiovascularandRenalDrugsAdvisoryCommittee/UCM220192.pdf.

Gallo, P., J. Chen, H. Quan, S. Menjoge, X. Luo, Y. Tanaka, G. Li, S. P. Ouyang, B. Binkowitz, E. Ibia, S. Talerico, and K. Ikeda (2011). Consistency of treatment effect across regions in multiregional clinical trials, part 2: monitoring, reporting and interpretation. *Drug Information Journal* **45**: 603–608.

Hung, H. M. J., S. J. Wang, and R. T. O'Neill (2010). Consideration of regional difference in design and analysis of multi-regional trials. *Pharmaceutical Statistics* **9**: 173–178.

Ibia, E. O. and B. Binkowitz (2011). Proceedings of the DIA Workshop on Multiregional Clinical Trials, October 26-27, 2010. *Drug Information Journal* **45**: 391–403.

ICH (1998). Statistical Principles for Clinical Trials E9. http://www.ich.org/products/guidelines/efficacy/article/efficacy-guidelines.html.

Li, Z., C. Chuang-Stein, and C. Hoseyni (2007). The probability of observing negative subgroup results when the treatment effect is positive and homogeneous across all subgroups. *Drug Information Journal* **41**: 47–56.

Marschner, I. C. (2010). Regional differences in multinational clinical trials: anticipating chance variation. *Clinical Trials* **7**: 147–156.

Quan, H., M. Li, J. Chen, P. Gallo, B. Binkowitz, E. Ibia, Y. Tanaka, S. P. Ouyang, X. Luo, G. Li, S. Menjoge, S. Talerico, and K. Ikeda (2010). Assessment of consistency of treatment effects in multiregional clinical trials. *Drug Information Journal* **44**: 617–632.

R Development Core Team (2014). R: A Language and Environment for Statistical Computing. www.R-project.org. R Foundation for Statistical Computing: Vienna, Austria.

Rothwell, P. M. (2005). Subgroup analysis in randomised controlled trials: importance, indications and interpretation. *Lancet* **365**: 176–186.

Schou, I. M. (2014). subgroup: Methods for exploring treatment effect heterogeneity in subgroup analysis of clinical trials. R package version 1.1. http://CRAN.R-project.org/package=subgroup.

Schou, I. M. and I. C. Marschner (2013). Meta-analysis of clinical trials with early stopping: an investigation of potential bias. *Statistics in Medicine* **32**: 4859–4874.

Serebruany, V. L. (2010). Aspirin dose and ticagrelor benefit in PLATO: fact or fiction? *Cardiology* **117**: 280–283.

Wallentin, L. et al. (2009). Ticagrelor versus clopidogrel in patients with acute coronary syndromes. *New England Journal of Medicine* **361**: 1045–1057.

Wang, R., S. W. Lagakos, J. H. Ware, D. J. Hunter, and J. M. Drazen (2007). Statistics in medicine - reporting of subgroup analyses in clinical trials. *The New England Journal of Medicine* **357**.21: 2189–2194.

Wedel, H., D. DeMets, P. Deedwania, B. Fagerberg, S. Goldstein, S. Gottlieb, A. Hjalmarson, J. Kjekshus, F. Waagstein, and J. Wikstrand on behalf of the MERIT-HF Study Group (2001). Challenges of subgroup analyses in multinational clinical trials: experiences from the MERIT-HF trial. *American Heart Journal* **142**: 502–511.

Wittes, J. (2013). "Why is this subgroup different from all other subgroups? Thoughts on regional differences in randomized clinical trials". *Proceedings of the Fourth Seattle Symposium in Biostatistics: Clinical Trials*. Ed. by W. B. S. Fleming T. R. New York, USA: Springer: 95–115.

Chapter 4

Software for exploring treatment effect heterogeneity

Manuscript ready for submission:
Schou, I. M. (2016). subgroup: A package for exploring treatment effect heterogeneity in subgroup analysis.

Abstract

Heterogeneity of treatment effect across subgroups is often assessed in randomised clinical trials. A test of interaction is the accepted statistical method for evaluating heterogeneity. However, this test is often underpowered which can make interpretation of apparent differences between subgroups difficult. Various measures of expected treatment effect heterogeneity have been proposed under an assumption of homogeneity across subgroups which provide a useful supplement to a formal test of interaction. This paper presents how these measures are implemented in the R computing environment using the **subgroup** package.

Keywords: clinical trial; heterogeneity; interaction; multi-country study; subgroup analysis

4.1 Motivation

Assessment of heterogeneity of treatment effects across subgroups is often conducted in randomised clinical trials (RCTs). The accepted statistically valid approach for this evaluation requires a test of interaction. However, RCTs are seldom adequately powered for these tests of interaction. Thus, the potential for a type II error can lead to speculation when

treatment effects are seemingly different between subgroups. In light of this limitation, evaluating the extent and nature of variation that can arise under an assumption of treatment effect homogeneity across subgroups provides a greater understanding of the play of chance. While a formal test of interaction will remain the best measure of treatment effect homogeneity for inferential purposes, various non-inferential measures which provide supplementary information on the extent of chance variation have been proposed (Marschner, 2010; Schou and Marschner, 2015). The **subgroup** package for the R computing environment implements these measures which include the expectations of the order statistics of the subgroup-specific treatment effects, the density function of the range of the treatment differences between subgroups, and the number of subgroups favouring the control treatment (Schou, 2014).

4.2 R package

Understanding the extent to which treatment differences can arise as a result of chance is useful at the design stage of a clinical trial (Marschner, 2010). It helps researchers get an appreciation for the treatment effect variations that can arise as a result of chance. Furthermore, prospective documentation of this potential chance variation can also mitigate speculation of treatment effect heterogeneity that might arise at the analysis stage of a clinical trial. At the analysis stage, observed measures of treatment effect variation can be compared with the expected measures in a non-inferential manner to gain a better understanding of the extent of variation, as a supplement to formal heterogeneity testing (Schou and Marschner, 2015). The **subgroup** package flexibly implements these methods and allows the user to indicate if the assessment of treatment effect variation is to be conducted at the design phase or the analysis phase of a clinical trial.

The graphical and numerical outputs produced by the **subgroup** package have been described in detail in Schou and Marschner (2015). These include the expectations of the order statistics of the subgroup-specific treatment effects, the density function of the range of the treatment differences between subgroups, and the number of subgroups favouring the

control treatment. In addition, the overall treatment difference, either as input by the user, or as calculated as the inverse variance weighted average, will also be output. The default graphical output is a 2×2 plot which presents these numerical measures. The outputs will differ slightly depending on whether it is a prospective design stage assessment of potential chance variation, or an analysis stage assessment of expected versus observed variation. If it is the latter, the observed measures will also be included in the plots. Furthermore, the probability of observing a measure at least as extreme as the observed will also be included as a non-inferential quantification in the graphs of the density function of the range of the treatment differences between subgroups, and the number of subgroups favouring the control treatment.

The computation of these measures is based on order statistics theory requiring combinatorial evaluations. Consequently, the computation time in the heteroscedastic case increases exponentially as the number of subgroups R, increases, as discussed later in the paper. Indeed, the computation time taken when $R > 20$ is significant, thereby compromising the practicality of running this package on a standard desktop or laptop computer. Therefore, the **subgroup** package defaults to a simulation based approach when $R > 20$ in the heteroscedastic case, unless the user explicitly specifies that a theory based approach is to be implemented. In the special case when the treatment effects are homoscedastic, the combinatorial evaluations are greatly simplified, thereby making the theory based calculations straightforward. In this case the **subgroup** package defaults to a theory based approach irrespective of the size of R.

The symmetry of the order statistics around the median is exploited to increase efficiency in the computation of the theory based expectations of the order statistics. As such, only the $(R/2)+0.5$ order statistics are calculated when R is odd, and $(R/2)$ order statistics are calculated when R is even. The remainder are computed using the symmetry attribute.

4.3 Distributional assumptions

The **subgroup** package has one main function, `subgroup`. The critical inputs to the function `subgroup` are the subgroup-specific treatment effects and the subgroup-specific standard errors. The methodology used in the computations assumes that the treatment difference in subgroup r is normally distributed with mean μ and standard deviation σ_r. Therefore, this is directly applicable to continuous endpoints that are normally distributed. Under large sample normality assumptions this distributional property can be readily extended to binary and time-to-event responses either directly, or to the log transformed treatment effect measure as described in Table 4.1.

Table 4.1: Binary and time-to-event outcome specification for the **subgroup** package.

Endpoint	
Difference in proportions	
Treatment difference	$p_{1r} - p_{2r}$
Standard error	$\left(\frac{p_{1r}(1-p_{1r})}{n_{1r}} - \frac{p_{2r}(1-p_{2r})}{n_{2r}}\right)^{1/2}$
Log relative risk	
Treatment difference	$log(p_{1r}) - log(p_{2r})$
Standard error	$\left(\frac{1}{n_{1r}p_{1r}} - \frac{1}{n_{1r}} + \frac{1}{n_{2r}p_{2r}} - \frac{1}{n_{2r}}\right)^{1/2}$
Log odds ratio	
Treatment difference	$log(p_{1r}) + log(1-p_{2r}) - log(p_{2r}) - log(1-p_{1r})$
Standard error	$\left(\frac{1}{n_{1r}p_{1r}} + \frac{1}{n_{1r}(1-p_{1r})} + \frac{1}{n_{2r}p_{2r}} + \frac{1}{n_{2r}(1-p_{2r})}\right)^{1/2}$
Log hazard ratio	
Treatment difference	β_r
Standard error	$\left(\frac{4}{e_{1r}+e_{2r}}\right)^{1/2}$

The notations used in Table 4.1 are as follows:

- n_{ir}: The number of subjects randomised to treatment group i and subgroup r, where $r = 1,..,R$.
- p_{ir}: The proportion of events in treatment group i and subgroup r.
- β_r: Log hazard ratio in subgroup r.

- e_{ir}: Number of events in treatment group i and subgroup r. When the randomisation ratio is 1 : 1, the variance of the log hazard ratio can be estimated by 4 times the reciprocal of the total number of events (Quan et al., 2010).

At the design stage of the study, the treatment difference is assumed to be the same across all the subgroups. As such, the treatment difference that will be input into the **subgroup** package will be the same for all subgroups. Likewise, the standard errors input into the **subgroup** package will be the same across subgroups if the standard deviation of the individual patient endpoint is assumed to be homoscedastic and the subgroups are anticipated to be the same size. However, if the sample sizes in each subgroup is anticipated to be different, or the underlying distribution of the individual patient measures is anticipated to be heteroscedastic, the input standard errors will vary across subgroups. It is recommended that the overall treatment difference is supplied by the user if the assessment is being conducted at the design stage; this would typically be the value used in the sample size calculation for the trial. At the analysis stage of the the study, the input data will be the observed subgroup treatment differences and its associated subgroup sample standard error. These may be available from a subgroup analysis to the user, or can be computed by the user as described in Table 4.1. If an estimate of the overall treatment difference is available, it is recommended that this value be provided by the user. If not, the **subgroup** package will calculate this as a reciprocal variance weighted estimate from the subgroup treatment differences.

4.4 Input arguments and return values

The **subgroup** package is defined as a single function, `subgroup`, which encapsulates several subroutines. This section describes the input arguments and the return values of the function `subgroup`.

The input arguments to the `subgroup` function and its default values are as follows:

- `data`: Numeric matrix of dimension $R \times 2$ where R is the number of subgroups. The

first column of this matrix will contain the treatment differences and the second column will contain the standard errors of the treatment differences.

- `overall.diff`: Numeric argument indicating the overall treatment difference if available. The default is `NULL`, in which case the weighted average of the subgroup treatment differences, weighted by the reciprocal of the variance of the subgroup-specific treatment differences will be generated. It is recommended that the user provides the overall treatment difference anticipated in the trial if the assessment is being conducted at the design stage of the study.

- `force.theoretical`: Logical argument with default set to `TRUE` if $R \leq 20$ or if the standard errors of the subgroup treatment effects are homogeneous. Otherwise, it defaults to `FALSE`. If set to `TRUE`, theoretical computations are used regardless of the number of subgroups, R, and a warning message is given to advise the user that the processing time may be significant if the standard errors across the subgroups are heterogeneous and $R > 20$. An error message is given if both `force.theoretical` and `force.simulation` are set to `TRUE`, and the routine will stop execution.

- `force.simulation`: Logical argument with default set to `FALSE` if $R \leq 20$ or if the standard errors of the subgroup treatment effects are homogeneous. Otherwise, it defaults to `TRUE`. If set to `TRUE`, simulation based computations are conducted regardless of the number of subgroups, R. An error message is given if both `force.theoretical` and `force.simulation` are set to `TRUE`, and the routine will stop execution.

- `design`: Logical argument with default set to `FALSE`. Allows the user to indicate that the outputs are to be created for a study at its design stage. The resulting plots will not present any observed measures.

- `plots`: Logical argument with default set to `TRUE`.

The following numerical components are returned by the **subgroup** package:

- `expectations`: A matrix of dimension $R \times 4$. The first two columns present the treatment differences and the standard errors contained in the dataset `data` as submitted by the user. The third column contains the expected ordered treatment differences, and the fourth column the order number.

- `favourcontrol`: A matrix of dimension $(R+1) \times 2$, where R is the number of subgroups. The first column contains the number of subgroups favouring control. The second contains the probability of that event. Here, a treatment effect that is > 0 is considered to favour the control.

- `rangedensity`: A matrix with 2 columns. The first column contains the sample space for the range which takes on values > 0. The second column contains the probability density.

- `overalldiff`: A numeric variable which returns the input argument `overall.diff` if provided by the user or contains the weighted mean treatment difference as calculated within the `subgroup` routine.

The user can manipulate these results to produce plots, or can opt to use the default plots produced by the **subgroup** package, using the `plots` argument.

4.5 Example

MERIT-HF (Metoprolol CR/XL Randomised Intervention Trial in Congestive Heart Failure) was a randomised placebo-controlled multi-country clinical trial which investigated whether all-cause mortality in patients with decreased ejection fraction and symptoms of heart failure could be lowered with the use of metoprolol controlled release/extended release (CR/XL) once daily, in addition to standard therapy (MERIT-HF Study Group, 1999; Wedel et al., 2001). At the analysis stage, the study had enrolled 3991 patients from 14 countries. Although the overall hazard ratio for all-cause mortality favoured metoprolol (hazard ratio=0.66), a post-hoc analysis of the primary endpoint by the country of randomi-

sation resulted in a hazard ratio of 1.05 in the US, with a non-significant p-value for the test of interaction (0.22). Wedel et al. (2001) investigated whether this seemingly unfavourable treatment effect in the US could be explained by any differences in baseline characteristics and concluded that this difference was likely an artefact of chance.

Table 4.2: Number of patients randomised and number of all-cause mortality events in the MERIT-HF study.

Country	Metoprolol		Placebo	
	Randomised	Events	Randomised	Events
Belgium	68	3	66	13
Czech Republic	123	9	124	17
Denmark	141	11	150	11
Finland	20	0	14	2
Germany	252	19	247	31
Hungary	211	16	212	29
Iceland	19	2	22	2
Norway	97	6	105	11
Poland	102	8	102	8
Sweden	39	2	46	9
Switzerland	21	0	21	1
The Netherlands	278	14	270	25
United Kingdom	87	4	83	9
United States	532	51	539	49
All countries	1990	145	2001	217

The number of patients randomised to each treatment arm in each country and the number of all-cause mortality events are reproduced in Table 4.2 (Wedel et al., 2001). To avoid analytical complications resulting from zero events in the metoprolol arm in Finland and Switzerland, the authors combined these countries with neighbouring countries (Finland with Denmark and Switzerland with The Netherlands) into randomisation regions. Taking a similar approach of combining these countries, the data presented in Table 4.2 can be used to descriptively evaluate whether the observed treatment effects differ from what might have been expected by chance under an assumption of treatment effect homogeneity. The overall relative risk in this study was 0.672, that is, a log relative risk of -0.398.

The code provided in Figure 4.1 demonstrates how the **subgroup** package can be used to analyse the MERIT-HF data presented in Table 4.2. The resulting output is presented in

```
# MERIT-HF study with 12 randomisation regions resulting from
# combining Finland with Denmark, and Switzerland with The Netherlands.

# Load the library
library(subgroup)
# Number randomised to the metoprolol arm.
n1 <- c(68, 123, 161, 252, 211, 19, 97, 102, 39, 299, 87, 532)
# Number randomised to the placebo arm.
n2 <- c(66, 124, 164, 247, 212, 22, 105, 102, 46, 291, 83, 539)
# Events in the metoprolol arm.
e1 <- c(3, 9, 11, 19, 16, 2, 6, 8, 2, 14, 4, 51)
# Events in the placebo arm.
e2 <- c(13, 17, 13, 31, 29, 2, 11, 8, 9, 26, 9, 49)
# Log relative risk in each subgroup.
difference<- log((e1/n1)/(e2/n2))
# Refer Table 4.1 for the formula for the standard error of the
# log relative risk.
se <- sqrt((1/e1) - (1/n1) + (1/e2) - (1/n2))
# Create the dataset for use.
data<- cbind(difference, se)
# Run the subgroup package to produce theory based outputs.
# Plot presented in Figure 4.2.
result1<- subgroup(data=data, overall.diff=-0.398)

# Run the subgroup package to produce simulation based outputs.
# Plot presented in Figure 4.3.
result2<- subgroup(data=data, overall.diff=-0.398, force.simulation=TRUE)
```

Figure 4.1: R code to produce outputs in Figures 4.2 and 4.3 using the **subgroup** package.

Figures 4.2 and 4.3. Here, Figure 4.2 presents the output from a theory based calculation, and Figure 4.3 the output from a simulation based assessment. With regards to potential heterogeneity of treatment effects across the randomisation regions, the graphical presentations in Figures 4.2 and 4.3 suggest that although the US had a log relative risk that was in favour of placebo, the overall variation in the subgroup treatment effects is not inconsistent with what would have been expected under an assumption of homogeneity. With regards to a comparison of the theory based approach with the simulation based approach, it can be seen that the simulation based approach is closely comparable to the theory based approach. This provides some assurance that the **subgroup** package gives reliable estimates of the measures of interest for $R > 20$, when it defaults to using a simulation based approach.

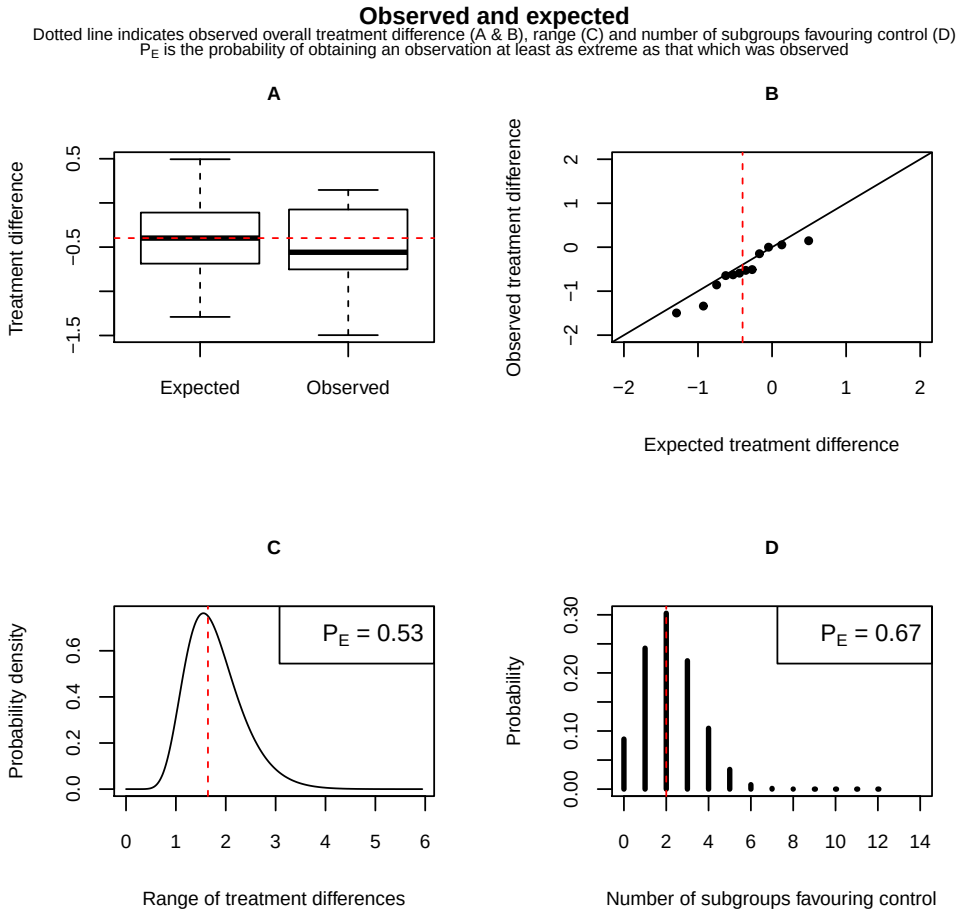

Figure 4.2: Theory based comparison of observed and expected treatment differences in a subgroup analysis of 12 randomisation regions in the MERIT-HF study.

The **subgroup** package can also be used at the design stage of a clinical trial. The MERIT-HF study had planned to randomise 1600 patients per treatment arm in a ratio of 1 : 1 across 14 countries (MERIT-HF Study Group, 1999). For the purposes of demonstrating the use of the **subgroup** package in the design context, suppose the planned number of events in the placebo group is 200, that is, an event rate of 12.5%, with a 30% risk reduction in those randomised to metoprolol, that is, a log relative risk of -0.357. In this example, the country of randomisation will define each subgroup. Therefore, the number randomised in each country will dictate the magnitude of the standard error associated with the log relative risk. This design stage evaluation is explored through the execution of the code presented in Figure 4.4 and resulted in the plot presented in Figure 4.5. Figure 4.5 suggests that we would expect to have around three countries with treatment effect estimates favouring the

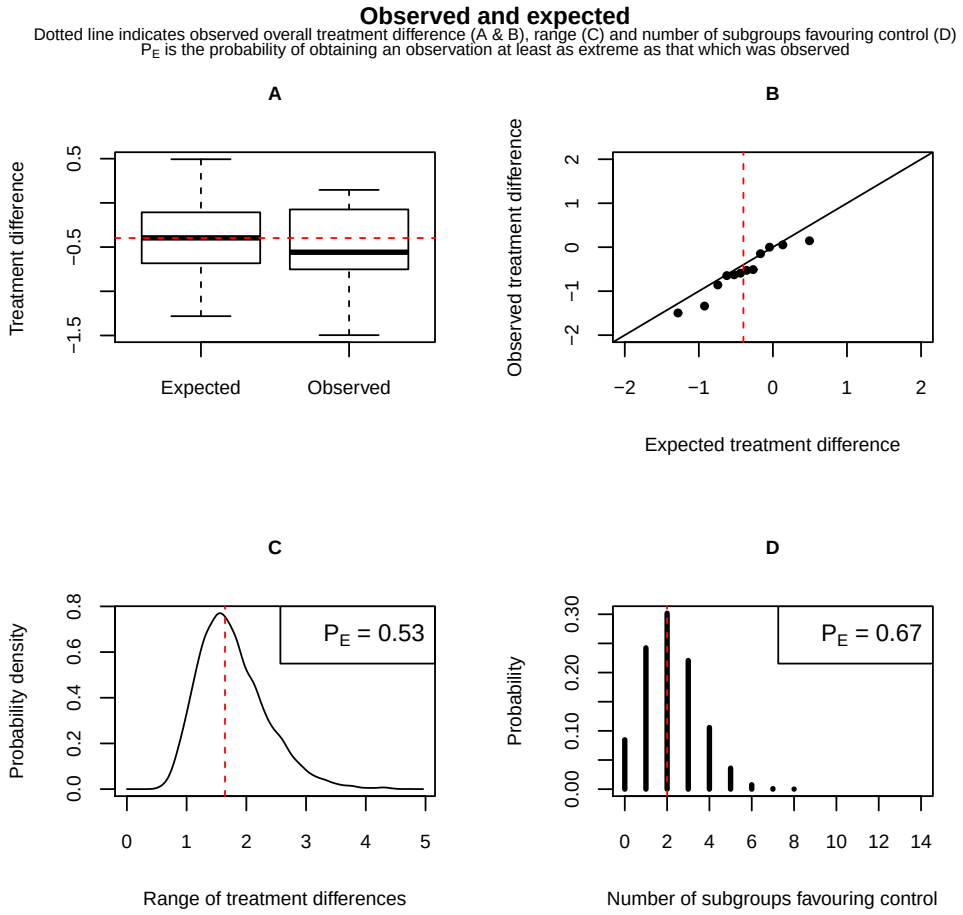

Figure 4.3: Simulation based comparison of observed and expected treatment differences in a subgroup analysis of 12 randomisation regions in the MERIT-HF study.

```
# Load the library
library(subgroup)
# Log relative risk in the 14 countries.
difference <- rep(log(0.7), 14)
# Patients per treatment arm in the 14 countries.
ni <- c(54, 99, 117, 14, 200, 170, 16, 81, 82, 34, 17, 220, 68, 429)
# Refer Table 4.1 for the formula for the standard error of the
# log relative risk.
se <- sqrt((1/(0.125*ni)) - (1/ni) + (1/(0.7*0.125*ni)) - (1/ni))
# Create the dataset for use.
data <- cbind(difference, se)
# Run the subgroup package.
result <- subgroup(data=data, design=TRUE, overall.diff=-0.357)
```

Figure 4.4: R code to produce output in Figure 4.5 using the **subgroup** package.

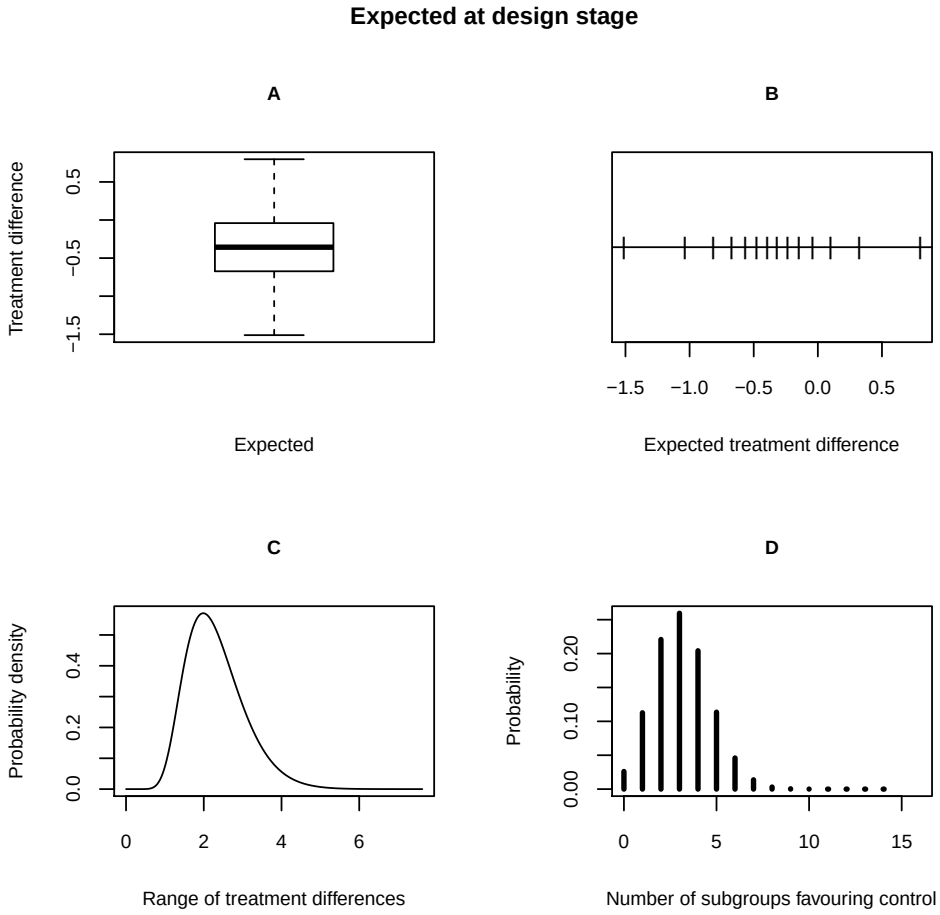

Figure 4.5: Expected variation in treatment effect in a subgroup analysis of 14 countries in the MERIT-HF study.

control, and that there is a high probability of this happening. In practice, this information might be used to manage expectations at the planning stage.

4.6 Computation time

Due to the combinatorial nature of the computations, the execution time of the **subgroup** package increases exponentially as the number of subgroups increases. Although the execution time depends greatly on the computer on which the **subgroup** package is implemented, as an illustrative example, the execution time of the **subgroup** package for subgroups ranging from $R = 2,..,15$ run on a 3.40 GHz core i7 processor is presented in Figure 4.6. This suggests that the user should consider the simulation approach which is a substantially faster

alternative to the theoretical approach in instances where the number of subgroups in an analysis is large.

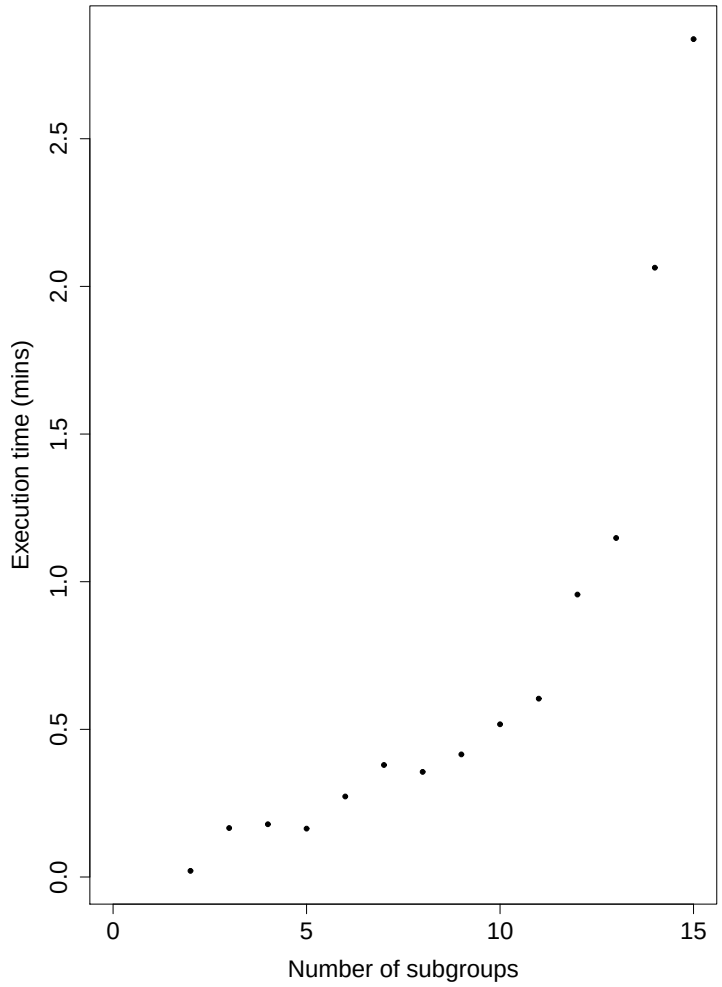

Figure 4.6: Execution time by number of subgroups.

4.7 Summary

This paper describes the context for the application of the **subgroup** package, the supplementary role it can play to a test of interaction in gaining an understanding of the nature and magnitude of chance treatment differences between subgroups, and provides examples of the execution of this package. The package can be used with different types of data; the example provided demonstrates its use in a binary context. Theoretical measures are desirable, but given the computational burden this poses with a large number of subgroups, the

simulation based approach provides a reliable substitute as demonstrated in the example. In conclusion, the **subgroup** package provides a suite of non-inferential tools that allows researchers to explore and understand the extent of subgroup treatment differences that could result from chance, and to compare these expected differences with those observed in a clinical trial.

4.8 References

Marschner, I. C. (2010). Regional differences in multinational clinical trials: anticipating chance variation. *Clinical Trials* **7**: 147–156.

MERIT-HF Study Group (1999). Effect of metoprolol CR/XL in chronic heart failure: metoprolol CR/XL randomised intervention trial in congestive heart failure (MERIT-HF). *Lancet* **353**: 2001–2007.

Quan, H., P. L. Zhao, J. Zhang, M. Roessner, and K. Aizawa (2010). Sample size considerations for Japanese patients in a multi-regional trial based on MHLW Guidance. *Pharmaceutical Statistics* **9**: 100–112.

Schou, I. M. (2014). subgroup: Methods for exploring treatment effect heterogeneity in subgroup analysis of clinical trials. R package version 1.1. http://CRAN.R-project.org/package=subgroup.

Schou, I. M. and I. C. Marschner (2015). Methods for exploring treatment effect heterogeneity in subgroup analysis: an application to global clinical trials. *Pharmaceutical Statistics* **14**: 44–55.

Wedel, H., D. DeMets, P. Deedwania, B. Fagerberg, S. Goldstein, S. Gottlieb, A. Hjalmarson, J. Kjekshus, F. Waagstein, and J. Wikstrand on behalf of the MERIT-HF Study Group

(2001). Challenges of subgroup analyses in multinational clinical trials: experiences from the MERIT-HF trial. *American Heart Journal* **142**: 502–511.

4.A Appendix: documentation for the R package

This appendix presents the documentation for the package **subgroup** available from the Comprehensive R Archive Network as:

Schou, I.M. subgroup: Methods for exploring treatment effect heterogeneity in subgroup analysis of clinical trials. R package version 1.1 2014. Available at: http://CRAN.R-project.org/package=subgroup.

This package has been developed as a single routine which implements the methods described in Chapters 3 and 4. The routine allows users to specify whether the evaluation is being conducted at the design stage or the analysis stage, and produces plots of the type presented in Figures 4.2, 4.3 and 4.5. The numerical output resulting from this routine can also be saved by the users to produce graphical presentations of their own choice. Furthermore, as the computation time can be quite substantial when the number of subgroups are large, the users also have the option to choose a simulation based output.

Package 'subgroup'

Type Package

Title Methods for exploring treatment effect heterogeneity in subgroup analysis of clinical trials

Version 1.1

Date 2014-07-31

Author I. Manjula Schou

Maintainer I. Manjula Schou <im.schou@yahoo.com.au>

Description Produces various measures of expected treatment effect heterogeneity under an assumption of homogeneity across subgroups. Graphical presentations are created to compare these expected differences with the observed differences.

License GPL-2 | GPL-3

Depends graphics, grDevices, utils, stats, R ($\geq$ 3.1)

`subgroup` *Compute measures for assessing treatment effect heterogeneity across subgroups in clinical trials and produce graphical presentations*

Description

This function produces various measures of expected treatment effect heterogeneity and allows graphical comparisons with the observed counterparts. The resulting measures can also be saved for creation of user-specified graphics.

Usage

```
subgroup(data, overall.diff = NA, force.theoretical = FALSE,
force.simulation = FALSE, design = FALSE, plots = TRUE)
```

Arguments

`data` A matrix of 2 columns containing the subgroup treatment effects and their associated standard errors. The treatment effects must be measured on a scale where 0 means no effect. For example, a mean difference, a log relative risk, log odds ratio, or log hazard ratio. Column 2 must contain the standard errors associated with the subgroup treatment effects in Column 1. Furthermore, a treatment effect that is > 0 is considered to favour the control.

`overall.diff` The overall treatment effect, provided optionally by the user. If not specified this is calculated within the `subgroup` routine as a reciprocal variance weighted mean.

`force.theoretical`
The default value for this argument is `FALSE`. If `TRUE`, theoretical computations are conducted regardless of the number of subgroups, `R`. If the standard errors of the subgroup treatment effects is heterogeneous and `R > 20`, a warning is given to advise the user that the processing time may be significant. An error message is given if both `force.theoretical` and `force.simulation` are set to `TRUE`, and the routine will stop execution.

`force.simulation`
The default value for this argument is `FALSE`. If `TRUE`, simulation based computations are conducted regardless of the number of subgroups, `R`. An error message is given if both `force.theoretical` and `force.simulation` are set to `TRUE`, and the routine will stop execution.

`design` The default value for this argument is `FALSE`. If `TRUE`, the plots will not include any observed measures.

`plots` The default value for this argument is `TRUE`. If `FALSE`, no plots will be displayed.

Details

Subgroup analysis principles require a significant test of interaction in order to claim heterogeneity of treatment effect. As clinical trials are typically underpowered for tests of interaction, overly optimistic expectations of treatment effect homogeneity can make interpretation difficult when treatment effects seemingly differ between subgroups. In addition to extending the ideas proposed by Marschner (2010), the package **subgroup** also implements some new measures, and provides a suite of graphical tools that allow visual comparison of the magnitude and nature of the observed and expected subgroup differences that can arise as an artefact of chance. These tools are intended to supplement a formal test of interaction in subgroup analyses, and are described in the manuscript Schou and Marschner (2014).

Three outputs are computed by the package. These include the following: the expectations of the ordered treatment effects, the probability density of the range, and the probability distribution of the number of subgroups favouring the control treatment. The user has the option to have the in-built plot suppressed. The content of the default plot

produced will depend on the user choice of an analysis stage evaluation or a design stage evaluation; if it is the former, the observed counterparts of the measures produced will be included in the plots.

Value

The following list of components is returned by the routine `subgroup`:

`expectations` A matrix of dimension `R x 4`, where `R` is the number of subgroups. The first two columns present the treatment differences and the standard errors contained in the dataset `data` as submitted by the user. The third column contains the expected ordered treatment differences, and the fourth column the order number.

`favourcontrol` A matrix of dimension `(R+1) x 2`, where `R` is the number of subgroups. The first column contains the number of subgroups favouring control. The second contains the probability of that event. Here, a treatment effect that is > 0 is considered to favour the control.

`rangedensity` A matrix with 2 columns. The first column contains the sample space for the range which takes on values > 0. The second column contains the probability density.

`overalldiff` A numeric variable which returns the input argument `overall.diff` if specified by the user or the reciprocal variance weighted mean treatment difference as calculated within the routine `subgroup`.

Author(s)

I. Manjula Schou

References

Marschner IC. Regional differences in multinational clinical trials: anticipating change variation. Clinical Trials 2010; 7:147-156.

Schou IM and Marschner IC. Methods for exploring treatment effect heterogeneity in subgroup analysis: an application to global clinical trials. Pharmaceutical Statistics 2015; 14: 44-55.

Examples

```
# Create dataset containing treatment differences ----------------------
# and standard errors. -------------------------------------------------
difference<-c(-0.163, -0.083, -0.030, 0.095)
se<-c(0.48, 0.27, 0.19, 0.39)
mydata<-cbind(difference, se)

# Example code to produce the expected measures together with the ------
# plot created by the subgroup routine for comparison against the ------
# observed differences. NOTE: The execution time increases as the ------
```

```
# number of subgroups increases. ---------------------------------------
test1<-subgroup(data=mydata)

# Expected measures produced by the subgroup routine that the user -----
# can manipulate to produce own graphics. ------------------------------
test1$overalldiff         # Overall difference between treatment groups.
test1$expectations        # Expectations of the ordered
                          # treatment differences.
test1$rangedensity[1:15,] # Sample of the probability density of
                          # the range.
test1$favourcontrol      # Prob dist of subgroups favouring the
                          # control treatment.

# Example code for evaluation of chance heterogeneity at the design -----
# stage. ---------------------------------------------------------------
test2<-subgroup(data=mydata, design=TRUE)

# Example code for simulation based evaluation of chance heterogeneity. -
test3<-subgroup(data=mydata, force.simulation=TRUE)
```

Chapter 5

Design of clinical trials with multiple hypothesis tests

Submitted article:

Schou, I. M. and I. C. Marschner (2016). Design of clinical trials involving multiple hypothesis tests with a common control. Submitted to *Biometrical Journal*.

Abstract

Randomised clinical trials comparing several treatments to a common control are often reported in the medical literature. For example, multiple experimental treatments may be compared with placebo, or in combination therapy trials, a combination therapy may be compared with each of its constituent mono-therapies. Such trials are typically designed using a balanced approach in which equal numbers of individuals are randomised to each arm, however, this can result in an inefficient use of resources. We provide a unified framework and new theoretical results for optimal design of such single-control multiple-comparator studies. We consider variance optimal designs based on D-, A- and E-optimality criteria, using a general model that allows for heteroscedasticity and a range of effect measures that include both continuous and binary outcomes. We demonstrate the sensitivity of these designs to the type of optimality criterion by showing that the optimal allocation ratios are systematically ordered according to the optimality criterion. Given this sensitivity to the optimality criterion, we argue that power optimality is a more suitable approach when designing clinical trials where testing is the objective. Weighted variance optimal designs are also discussed which, like power optimal designs, allow the treatment difference to play a

major role in determining allocation ratios. We illustrate our methods using two real clinical trial examples taken from the medical literature. General guidelines on the use of optimal designs in single-control multiple-comparator trials are also provided.

Keywords: complete power; Dunnett adjustment; minimal power; multiple testing; optimal design; weighted optimality

5.1 Introduction

Clinical trials involving comparisons of several treatments with a common control arise quite frequently in the medical literature. For example, in some trials there are multiple experimental treatments that are each compared with placebo. Combination therapy studies often have a similar design, in which each of the constituent mono-therapies is compared with the combination therapy. Although it is well-known that unbalanced allocation is more efficient in experiments that compare several treatments with a common control (Fleiss, 1986), balanced designs are usually the preferred approach in clinical trials.

In the classical experimental design literature, unbalanced allocation based on the optimisation of a variance measure is well established (Atkinson and Donev, 1992; Hedayat, Jacroux, and Majumdar, 1988). Such considerations lead to standard design criteria such as D-, A- and E-optimality. Weighted versions of these variance-based criteria have also been considered in the literature, allowing for the possibility that different comparisons have different importance (Morgan and Wang, 2010). Over the years various authors have discussed the application of variance optimality criteria in the specific context of clinical trials (Zhu and Wong, 2000; Wong and Zhu, 2008) for the purpose of producing designs that are more efficient than the standard balanced design. Other types of optimality criteria have also been discussed for clinical trials, particularly power optimality (Marschner, 2007). In this paper we will provide a general and unified discussion of these concepts in the context of clinical trials comparing several treatments with a common control. Our main goal will be to compare the various approaches to optimal design with a view to producing some general

guidelines on their use in clinical trials.

The first part of our paper will present new results which unify the optimal designs under different types of variance optimality, both unweighted and weighted, and will demonstrate the sensitivity of the optimal designs to the chosen optimality criterion. These results will be presented in the context of a general model allowing for heteroscedasticity and a range of effect measures that include both continuous and binary outcomes. In view of the sensitivity of the design to the chosen optimality criterion, we argue that optimisation of power is usually more appropriate for clinical trials, where the focus is typically hypothesis testing rather than estimation. The second part of our paper therefore focuses on exploring the relationships between variance optimal designs and power optimal designs. Since power optimal designs are generally more complex than variance optimal designs, we provide some numerical results supporting the approximation of power optimal design allocation ratios using appropriately chosen variance optimal design allocation ratios. Examples based on the design of published clinical trials are used to demonstrate the application of our results in practice. Finally, some general guidelines on the use of optimal designs in single-control multiple-comparator trials will be provided.

5.2 Assumptions

5.2.1 General model

We consider a clinical trial in which $k \geq 2$ groups are each compared with a single control group, and we suppose that these are the only comparisons undertaken. Our assumptions will cover a wide range of standard difference measures and will allow for heteroscedasticity and unequal effect sizes. We will use the term control for the common comparator group, although our assumptions also cover the situation in which the common comparator is actually an experimental group, such as a design comparing a combination treatment with each of its constituent treatments.

Let $\hat{\Delta}_i = T_i - T_0$ be the difference measure for comparing group $i = 1, \ldots, k$ with the control

group $i = 0$. It is assumed that $T_i, i = 0, \ldots, k$, are independent with $\mathrm{N}(\theta_i, v_i/n_i)$ distributions where n_i is the number of observations in group i and the total number of observations is $N = n_0 + \cdots + n_k$. The variance term v_i may depend on the group mean θ_i and a group-specific scale parameter σ_i through $v_i = v(\theta_i, \sigma_i)$ for some function v. Letting $\gamma_i = n_i/N$ be the proportion of the total sample size that is allocated to group i, we have that $\hat{\Delta} = (\hat{\Delta}_1, \ldots, \hat{\Delta}_k)$ is multivariate normal with mean vector $\delta = (\delta_1, \ldots, \delta_k)$ where $\delta_i = \theta_i - \theta_0$. The variance-covariance matrix of $\hat{\Delta}$ is Σ/N, where Σ is the $k \times k$ matrix with all off-diagonals equal to v_0/γ_0 and i^{th} diagonal element equal to $v_0/\gamma_0 + v_i/\gamma_i$. That is,

$$\Sigma = \mathrm{diag}\left(\frac{v_1}{\gamma_1}, \ldots, \frac{v_k}{\gamma_k}\right) + \frac{v_0}{\gamma_0}\mathbf{1}_k\mathbf{1}_k{}^{\mathrm{T}} \tag{5.1}$$

where $\mathbf{1}_k$ is the $k \times 1$ matrix with all elements equal to 1. The group-specific standard deviation ratios defined as

$$r_i = \sqrt{\frac{v_i}{v_0}} \qquad i = 1, \ldots, k$$

quantify the heteroscedasticity relative to the control group and will be important in subsequent sections.

The above assumptions allow for a very general range of clinical trial designs. The most basic setting is that of a continuous outcome with heteroscedastic normal distribution where the difference is measured using a difference in the group-specific means θ_i, and a variance function that is independent of the mean, $v(\theta_i, \sigma_i) = \sigma_i^2$. However, the assumptions also allow for other types of outcomes where large sample normality of the summary measure holds for each group. This includes a wide range of commonly used difference measures, including the standard measures for independent binary outcomes. Clustered binary outcomes can also be accommodated, with an extra-binomial variance factor of $\sigma_i^2 = \sigma^2$ within each cluster, as can survival outcomes with constant hazard rate. Some of the experimental designs accommodated under the general model are illustrated in Table 5.1, and designs based on continuous and binary outcomes will be used for numerical illustration throughout the paper. The $v(\theta_i, \sigma_i) = 1$ applied to the survival outcome in Table 5.1 requires a special com-

Table 5.1: Examples of outcomes and difference measures covered by the assumptions.

Outcome type	θ_i	Difference measure	$v(\theta_i, \sigma_i)$
continuous	mean response	mean difference	σ_i^2
binary	risk	risk difference	$\theta_i(1-\theta_i)$
binary	log risk	relative risk (log scale)	$e^{-\theta_i}-1$
binary	log odds	odds ratio (log scale)	$e^{-\theta_i}(1+e^{\theta_i})^2$
clustered binary	risk	risk difference	$\sigma^2\theta_i(1-\theta_i)$
survival	log hazard	hazard ratio (log scale)	1

ment. Assuming exponential survival with different hazards across the treatment groups, the log of the estimated hazard ratio is approximately normally distributed with an expectation of the log hazard ratio and a variance of $(1/n_0)+(1/n_i)$ (George and Desu, 1974). Consequently, substituting $n_i = N\gamma_i$ results in $v(\theta_i, \sigma_i) = 1$.

For a given difference measure and an assumption about the parameter values, the group-specific sample sizes can then be determined in such a way that the design is optimal in some sense. There are many approaches to defining optimality, and in the next section we review the three main approaches based on minimisation of variance. We will present new results that unify the form of the optimal designs under these criteria, which then allows us to compare the different approaches and illustrate the sensitivity of the optimal design to the chosen optimality criterion. We will argue that these criteria are generally inappropriate for clinical trials involving hypothesis tests concerning group differences. This will motivate the investigation of other optimality criteria that are more appropriate for studies aimed at testing for group differences, and we will lead to some general recommendations on their use.

5.2.2 Special cases

Before discussing optimal designs arising from the general model, we note that the general model has a number of special cases that will be useful in the theoretical and numerical investigations undertaken in subsequent sections. Firstly, we define four sub-models of the general model, having increasingly general assumptions concerning the standard deviation

ratios r_i for $i = 1, \ldots, k$. These sub-models are:

(i) *homoscedasticity*: equal variances in all groups, that is $r_i = 1$

(ii) *constant heteroscedasticity*: equal variances in all non-control groups, that is $r_i = r$

(iii) *consistent heteroscedasticity*: non-control variances are either all greater or all less than the control variance, that is $\max(r_1, \ldots, r_k) \leq 1$ or $\min(r_1, \ldots, r_k) \geq 1$

(iv) *general heteroscedasticity*: no restrictions on the relative magnitudes of the standard deviation ratios.

In practice the homoscedastic model is likely to be the most commonly used model for continuous outcomes, while the model with constant heteroscedasticity is likely to be the most commonly used model for other types of outcomes. In particular, observe that the model with constant heteroscedasticity would be appropriate for binary outcomes with common effect size and variance functions described in Table 5.1.

In addition to the four sub-models described above, there are two special cases that will be useful in our numerical investigations, both of which have $k = 2$ comparisons between 3 groups. The first model is a special case of the model with consistent heteroscedasticity, in which $r_1 = 1$ and $r_2 = r$. We will refer to this model as the 3-group consistent heteroscedasticity model. The second model is a special case of the general model, in which $r_1 = r$ and $r_2 = 1/r$. We will refer to this model as the 3-group general model. Both of these special cases are dependent on a single quantity r that determines the extent of heteroscedasticity, which makes the models convenient for displaying the dependence of the optimal designs on the nature of the heteroscedasticity.

5.3 Variance optimality

In the experimental design literature the most common approaches to optimal design are based on variance optimality criteria, which involve minimizing some measure of the sam-

pling variation exhibited by $\hat{\Delta}$. That is, for a given overall sample size N, variance optimality corresponds to choosing the allocation proportions γ_i in such a way that Σ is minimized in some sense. This leads to an optimal design that is independent of the overall sample size N, and allocates the N observations in optimal proportions across the groups.

There are many senses in which Σ can be minimized. In this paper we focus on three of the most popular approaches: (i) D-optimality, which corresponds to minimisation of the $\det(\Sigma)$; (ii) A-optimality, which corresponds to minimisation of the $\text{trace}(\Sigma)$; and (iii) E-optimality, which corresponds to minimisation of the maximum eigenvalue of Σ. In the present context, it turns out that the three variance optimality criteria can be unified by a common form for the optimal design, as presented below. We defer a justification of this form until Section 5, where it arises as a special case of more general results on weighted optimality criteria. In the meantime we make use of this unified form to facilitate comparisons between the different approaches to variance optimality.

5.3.1 Unified form

A unified form for the optimal designs arising from the three variance optimality criteria can be expressed as a function of the number of comparisons k, and the group-specific standard deviation ratios r_i. The following proposition specifies the optimal group allocation proportions, and is in fact a special case of a more general result presented in Section 5.5 for weighted optimality criteria.

Proposition 5.3.1 *For the D-, A- and E-optimality criteria, the optimal design for the model with general heteroscedasticity is*

$$\gamma_0 = \frac{1}{1+\overline{R}} \qquad \textit{and} \qquad \gamma_i = \gamma_0 u_i R(x_i, y_i, z_i) \qquad i = 1, \ldots, k$$

where $R(x,y,z) = [x + y(z-x)]^{-1}$ *and* (u_i, x_i, y_i, z_i) *depends on the chosen optimality crite-*

rion according to Table 5.2, with $\overline{R}$ being the weighted average

$$\overline{R} = \sum_{i=1}^{k} u_i R(x_i, y_i, z_i).$$

Table 5.2: Quantities used in the unified form of the optimal designs for D-, A- and E-optimality.

Criterion	x_i	y_i	z_i	u_i	$\hat{u}_i$
D	γ_0	$\frac{1}{r_i^2}$	$\frac{1}{k}$	$\frac{1}{k}$	$\frac{1}{k}$
A	0	$\frac{1}{r_i}$	1	$\frac{1}{\sqrt{k}}$	$\frac{w_i}{\sqrt{\sum_{j=1}^{k} w_j^2}}$
E	0	$\frac{1}{r_i}$	1	$\frac{(1+r_i)}{\sum_{j=1}^{k}(1+r_j)}$	$\frac{w_i^2(1+r_i)}{\sum_{j=1}^{k} w_j^2(1+r_j)}$

The proof of Proposition 5.3.1 is provided in the Web Appendix as a special case of the optimal designs from the weighted optimality criteria discussed in Section 5.5. The form of the D- and A-optimal designs in Proposition 5.3.1 reduce to the forms that are equivalent to those given previously; see for example Wong and Zhu (2008). On the other hand the explicit E-optimal design, and the weighted version of Proposition 5.3.1 given in Section 5.5, are both new. Note that since $x_i = \gamma_0$ for the D-optimality criterion, the D-optimal design for the general model requires a numerical solution except in some very simple special cases. On the other hand, the A- and E-optimal designs are always explicit. Also note that not all quantities specified in Table 5.2 are required for Proposition 5.3.1, in particular, the quantities $\hat{u}_i$ will not be used until Section 5.5 when our discussion is extended to weighted versions of the optimality criteria.

The unified form specified in Proposition 5.3.1 shows that the optimal design for each of the optimality criteria is governed by the weighted average $\overline{R}$ of the quantities $R(x_i, y_i, z_i)$. This connection between the three criteria allows some theoretical comparisons to be made between the corresponding optimal designs. These comparisons are useful for highlighting

systematic differences between the optimality criteria, and demonstrating the sensitivity of the optimal design to the chosen optimality criterion. In Sections 5.3.2 and 5.3.3 we review these theoretical comparisons for various models discussed in Section 5.2, along with some numerical comparisons. In Section 5.3.4 we discuss the implications of these comparisons.

5.3.2 Homoscedasticity and constant heteroscedasticity

An obvious consequence of Proposition 5.3.1 is that the three optimality criteria differ in the way they allocate sample size to the control group. Less obvious is the fact that for large sub-classes of the general model, these differences lead to systematic orderings of the three criteria with respect to their propensity to allocate sample size to the control and non-control groups. We begin by considering this for the model with homoscedasticity or constant heteroscedasticity.

Let $\gamma_0^{(D)}$, $\gamma_0^{(A)}$ and $\gamma_0^{(E)}$ be the optimal control proportions under D-, A- and E-optimality, respectively. Then the following proposition summarises a general ordering of the three optimal proportions.

Proposition 5.3.2 *For the model with constant heteroscedasticity, that is with $r_i = r$ for $i = 1,\ldots,k$, including the homoscedastic model with $r = 1$, the optimal control group proportions satisfy*

$$\gamma_0^{(D)} < \gamma_0^{(A)} < \gamma_0^{(E)}.$$

The inequality $\gamma_0^{(A)} < \gamma_0^{(E)}$ follows immediately from Proposition 5.3.1 which yields

$$\gamma_0^{(A)} = \frac{1}{1+r\sqrt{k}} \qquad \text{and} \qquad \gamma_0^{(E)} = \frac{1}{1+r} \qquad \text{when } r_i = r$$

while the proof of $\gamma_0^{(D)} < \gamma_0^{(A)}$ is more involved and is provided in the Web Appendix.

Proposition 5.3.2 shows that, for an important class of models including those most likely to be used in practice, the E-optimality criterion places greatest emphasis on allocating sample

size to the common control, whereas the D-optimality criterion places least emphasis on the common control and the A-optimality criterion is intermediate to the other two. Numerical comparisons of the three approaches reveal that the theoretical tendencies reflected in Proposition 5.3.2 can involve very large differences between the optimal designs in practice. This is most easily illustrated for the homoscedastic model with $r_i = 1$. In this case the E-optimal design always allocates half the available sample size to the control group. In contrast, the D-optimal design is always the balanced design, while the A-optimal design is the well-known "root-k" design (Fleiss, 1986; Dunnett, 1955). This leads to the homoscedastic special case of Proposition 5.3.2

$$\gamma_0^{(D)} = \frac{1}{1+k} \quad < \quad \gamma_0^{(A)} = \frac{1}{1+\sqrt{k}} \quad < \quad \gamma_0^{(E)} = \frac{1}{2} \qquad \text{when } r_i = 1.$$

Thus, for example, with $k = 4$ comparisons the optimal allocations to the control group will range from 20% to 50%, depending on the optimality criterion.

Such large differences are not limited to the homoscedastic model. Even larger differences are possible with constant heteroscedasticity. Figure 5.1 displays the potentially wide variation in optimal designs from the model with constant heteroscedasticity, using a range of values for k and the common standard deviation ratio $r_i = r$. It can be seen that when the common control has greater variance than the other groups, that is when $r < 1$, the differences between the three optimality criteria are exacerbated relative to the homoscedastic model. Thus, for example, in Panel B of Figure 5.1 it can be seen that with $k = 4$ comparisons and $r = 0.25$, the optimal allocations to the control group range from approximately 25% to 80%, depending on the optimality criterion.

5.3.3 Consistent and general heteroscedasticity

For the model with general heteroscedasticity, the ordering in Proposition 5.3.2 no longer holds in general. A counterexample is provided by the 3-group general model described in Section 5.2.2, in which $k = 2$ and $r_1 = 1/r_2 = r$. Panel B of Figure 5.2 displays the optimal

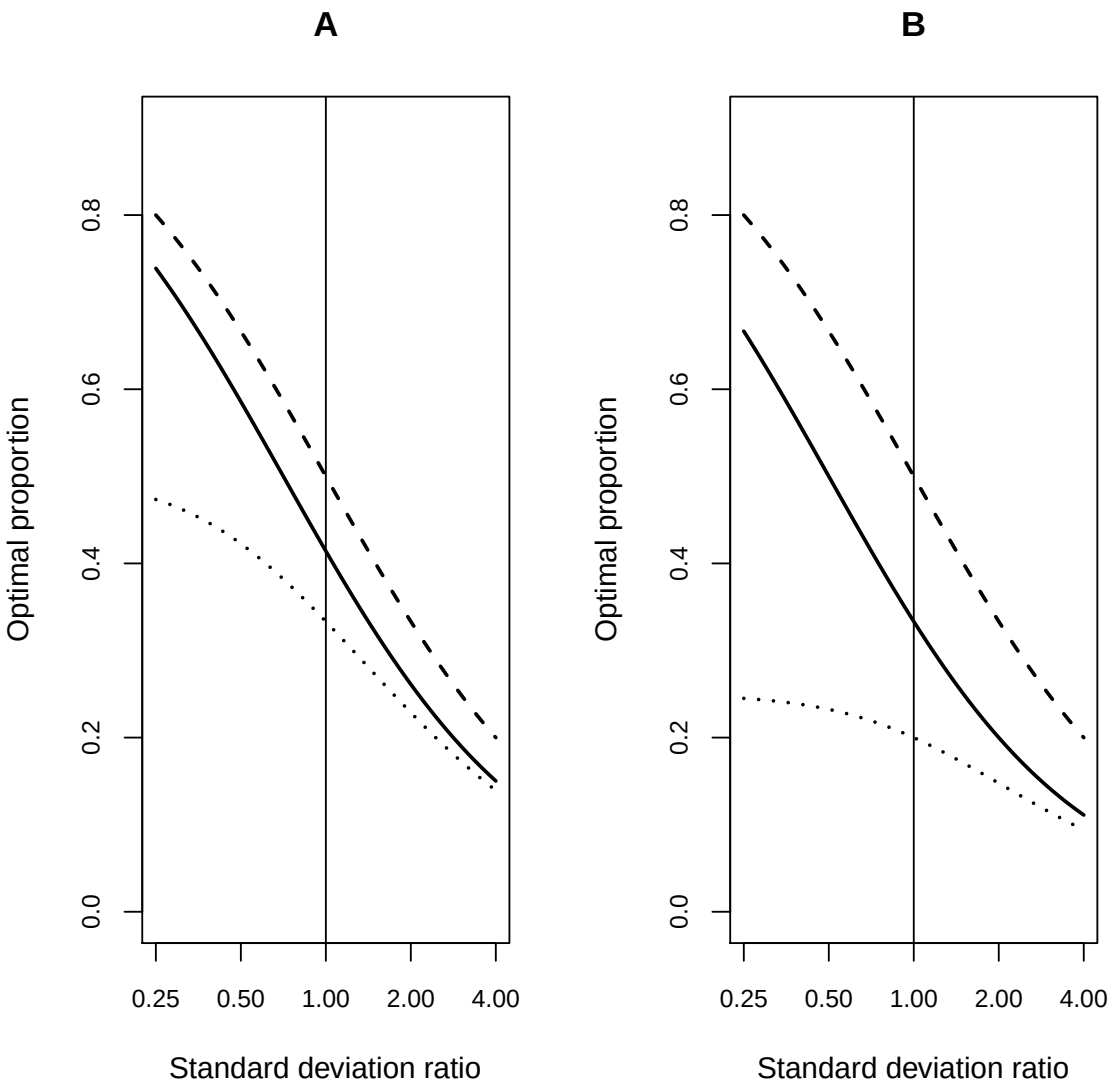

Figure 5.1: Optimal control group proportions (γ_0) as a function of the standard deviation ratio (r) in the constant heteroscedasticity model with E-optimality (dashed line), A-optimality (solid line) and D-optimality (dotted line). The two panels correspond to $k = 2$ (Panel A) or $k = 4$ (Panel B).

proportion allocated to the common control group for a range of values of r. It can be seen that no general ordering exists between the optimal control proportions from the three criteria. Nonetheless, this counterexample does show that even when there is no systematic ordering between the optimality criteria, fundamental differences can exist. In particular, for the limiting cases in which $r \to \infty$ or $r \to 0$, the D-optimal design allocates 50% of the sample size to the common control group, whereas the other two criteria allocate 0%. Furthermore, the relationship between the optimal proportion and the standard deviation ratio r for the D-optimal design is the reverse of the corresponding relationship for the other

two criteria. This exposes quite fundamental sensitivity of the optimal design to the chosen optimality criterion.

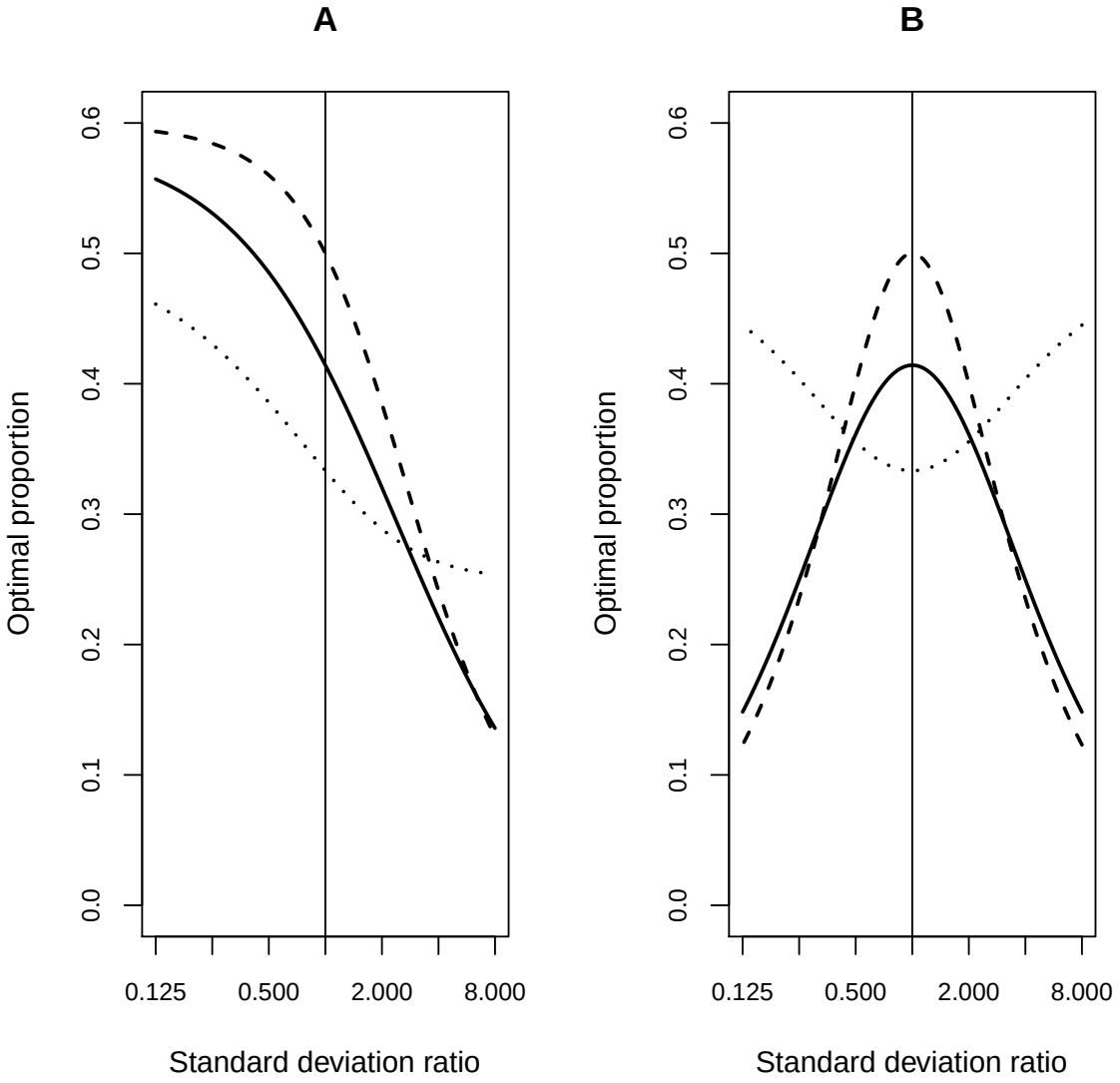

Figure 5.2: Optimal control group proportions (γ_0) as a function of the standard deviation ratio (r) for the 3-group special cases of consistent heteroscedasticity where $r_1 = 1$ and $r_2 = r$ (Panel A), and general heteroscedasticity where $r_1 = r$ and $r_2 = 1/r$ (Panel B), with E-optimality (dashed line), A-optimality (solid line) and D-optimality (dotted line).

Although Proposition 5.3.2 does not hold in general, there are some systematic differences between the optimality criteria that continue to hold under non-constant heteroscedasticity. In particular, consider the model with consistent heteroscedasticity in which the common control exhibits more variability than the other groups. In this case, all three criteria will favour the control over the non-control groups, but the D-optimality criterion has a tendency to favour the control to a lesser extent. This is reflected in the following result.

Proposition 5.3.3 *For the model with consistent heteroscedasticity where the common control exhibits most variability, that is* $\max(r_1,\ldots,r_k) \leq 1$*, the optimal control group proportions satisfy*

$$\gamma_0^{(D)} < \gamma_0^{(A)} < \gamma_0^{(E)}.$$

The ordering in Proposition 5.3.3 does not hold for consistent heteroscedasticity in which the common control exhibits less variability than the other groups. Counterexamples for these situations are provided by the 3-group consistent heteroscedasticity model described in Section 5.2.2 in which $k = 2$, $r_1 = 1$ and $r_2 = r$. Panel A of Figure 5.2 displays the optimal proportion allocated to the common control group for a range of values of r. It can be seen that the ordering presented in Proposition 5.3.3 is reflected in Panel A when $r \leq 1$. However, when $r \geq 1$ there is no general ordering between any of the three criteria.

Additionally, there is a systematic ordering of the optimal non-control proportions in the A- and E-optimal designs. Under the fully general heteroscedasticity model, the E-optimality criterion has a tendency to shift observations into the more variable groups to a greater extent than the A-optimality criterion. To examine this, define the following allocation ratios of the non-control groups

$$g_{ij} = \frac{\gamma_i}{\gamma_j} \qquad i,j = 1,\ldots,k$$

and let $g_{ij}^{(A)}$ and $g_{ij}^{(E)}$ be g_{ij} under the A- and E-optimality criteria, respectively. Then the following result shows that the E-optimality criterion tends to favour the more variable groups to a greater extent than the A-optimality criterion.

Proposition 5.3.4 *For the model with general heteroscedasticity, if* $r_i \geq r_j$ *then*

$$1 \leq g_{ij}^{(A)} \leq g_{ij}^{(E)} \qquad i,j = 1,\ldots,k.$$

It is obvious that g_{ij} should be no less than 1 under both optimality criteria, since a more variable group should receive more sample size than a less variable group. However, Propo-

sition 5.3.4 additionally says that the relative increase in sample size allocated to the more variable group is greater under the E-optimality criterion than under the A-optimality criterion. The proofs of Propositions 5.3.3 and 5.3.4, which stem from the unified form of the optimal designs in Proposition 5.3.1, are provided in the Web Appendix.

5.3.4 Implications

The theoretical and numerical results of Sections 5.3.2 and 5.3.3 illustrate the potential for systematic and large differences in the optimal designs associated with the three optimality criteria. A schematic summary of the relationships between the optimal proportions allocated to the common control under the various model assumptions is presented in Figure 5.3. While the orderings of the optimality criteria are not uniform across all models, broadly speaking Figure 5.3 shows that the E- and A-optimality criteria tend to allocate more sample size to the control group than the D-optimality criteria, while the numerical results of Sections 5.3.2 and 5.3.3 show that the differences can be substantial.

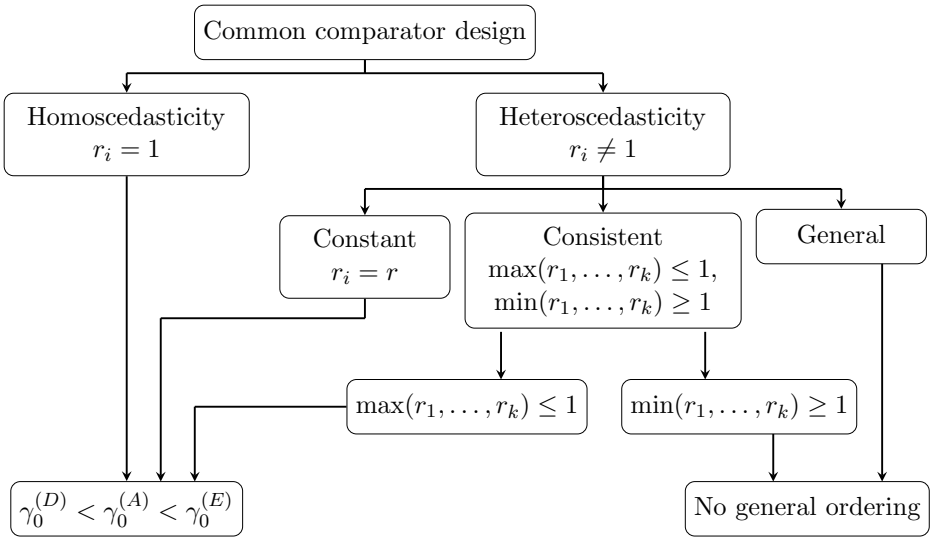

Figure 5.3: Schematic summary of the relationships between the optimal proportions allocated to the common control.

The sensitivity of the optimal design to the chosen optimality criteria means that it is important to link the objectives of an experiment with the interpretation of the optimality criterion

being used. While the three optimality criteria are defined using measures of the magnitude of the matrix Σ, they each have statistical interpretations that could potentially facilitate such linkage. In the case of *D*- and *A*-optimality these interpretations are well-known to correspond, respectively, to minimisation of the volume of the confidence ellipsoid for δ and minimisation of the average of the k effect variances, $\text{Var}(\hat{\Delta}_i)$ (Atkinson and Donev, 1992). The statistical interpretation of *E*-optimality is more complex, but can be summarized as the minimisation of the largest variance over all contrasts of θ_0 with a normalized average of the other θ_i (Morgan and Wang, 2010).

While it may be possible to link one of these interpretations to the objectives of the experiment in some contexts, in practice this is often not the case. Such linkage is particularly problematic for experiments involving the testing of hypotheses concerning treatment effects, such as clinical trials. These experiments are typically designed to achieve adequate power for testing hypotheses of no treatment difference, rather than to achieve adequate sampling variation for estimation purposes. In practice it may not be clear how the various criteria for minimizing sampling variance relate to the power of the experiment.

A further problem with variance optimality is that it only provides rules for optimally allocating a fixed overall sample size N among the $k+1$ groups. In practice there remains the task of choosing N. One approach would be to choose N so that δ is estimated with an acceptable level of precision, as determined by the optimality criterion. For example, if *A*-optimality is being used, then N could be determined so as to achieve an acceptable value for $\text{trace}(\Sigma)$. However, such an approach is uncommon in many applications, particularly clinical trials. In many contexts N is more likely to be determined with a view to achieving acceptable power for testing hypotheses about the group differences. This raises the possibility of using power to determine the optimal allocation proportions, which leads to an alternative approach to optimal design.

5.4 Power optimality

Variance optimality tackles the experimental design from an estimation perspective. In particular, it assumes that the primary objective is to estimate δ as precisely as possible regardless of its magnitude. This means that optimal designs based on variance optimality criteria do not depend on δ. In many contexts, particularly clinical trials, the primary objective is to assess whether the non-control groups differ from the control using hypothesis testing, in which case power is more important than sampling variability. This leads to the notion of power optimality, in which the optimal design is determined by maximizing the power for testing hypotheses concerning δ. Although the two notions of optimality are related, a key distinction is that a larger δ_i does not need to be estimated as precisely as a smaller δ_i when power is the primary focus. Optimality criteria based on power therefore lead to optimal designs that depend on δ by implicitly down-weighting comparisons having larger δ_i. As with variance optimality criteria, there is more than one way in which power can be optimized, leading to multiple versions of power optimality resulting from the way the hypotheses are expressed.

Consider the testing situation in which $H_{0i} : \theta_i - \theta_0 = 0$ represents the null hypothesis that defines the comparison of the i^{th} treatment group with the common control. In the current context, we consider two versions of power, which in the terminology of Westfall et al. (1999), are referred to as complete power and minimal power. Complete power can be interpreted as the probability of rejecting *all* H_{0i} while minimal power is the probability of rejecting *at least one* H_{0i}. Of note is the need to control the type I error for multiple testing in the minimal power design, whereas controlling for multiple tests is not necessary in the complete power design. We now consider each of these versions of power from the point of view of optimal design.

5.4.1 Complete and minimal power

Let Z_i represent the test statistic associated with each null hypothesis H_{0i} and is given by

$$Z_i = \sqrt{N}\hat{\Delta}_i \left(\frac{v_0}{\gamma_0} + \frac{v_i}{\gamma_i} \right)^{-\frac{1}{2}}$$

for $i = 1, \ldots, k$. Under the null hypothesis the collection of these k test statistics is multivariate normal with a zero mean, unit variance, and a correlation between Z_i and Z_j given by

$$\frac{v_0}{\gamma_0} \left\{ \left(\frac{v_0}{\gamma_0} + \frac{v_i}{\gamma_i} \right) \left(\frac{v_0}{\gamma_0} + \frac{v_j}{\gamma_j} \right) \right\}^{-\frac{1}{2}}.$$

As mentioned earlier, complete power is the probability that *all* of these Z_i are sufficiently extreme. In the context of one-sided hypothesis testing this can be defined as exceeding the critical value $\Phi^{-1}(1-\alpha)$ where α represents the significance level, whereas for two-sided testing it can be defined as exceeding $\Phi^{-1}(1-\alpha/2)$ in absolute value. Denoting the multivariate cumulative distribution function associated with $Z_1, \ldots, Z_k$ by $\Phi_{k,\gamma}$, complete power in the one-sided case can be written as

$$P_C(N) = \Phi_{k,\gamma}\left(z_i - \Phi^{-1}(1-\alpha), \ldots, z_k - \Phi^{-1}(1-\alpha) \right) \tag{5.2}$$

where $\gamma = (\gamma_0, \ldots, \gamma_k)$ and $z_i = \sqrt{N}\delta_i \left(\frac{v_0}{\gamma_0} + \frac{v_i}{\gamma_i} \right)^{-\frac{1}{2}}$ (Marschner, 2007). An analogous expression can also be provided for the two-sided case.

Minimal power requires that at least one of the Z_i is sufficiently large, or conversely that none of the Z_i exceed the critical value. In the one-sided case this can be written as

$$P_M(N) = 1 - \Phi_{k,\gamma}\left(\Phi^{-1}(1 - A(\alpha,k,\gamma)) - z_1, \ldots, \Phi^{-1}(1 - A(\alpha,k,\gamma)) - z_k \right) \tag{5.3}$$

where $A(\alpha,k,\gamma)$ denotes a pair-wise comparison significance level that results in a family-wise significance level of α (Marschner, 2007). For example, if a Bonferroni adjustment is used, $A(\alpha,k,\gamma) = \alpha/k$, which does not depend on γ. On the other hand if a Dunnett

adjustment is used, $A(\alpha,k,\gamma)$ does depend on γ (Dunnett, 1955; Marschner, 2007). The Dunnett adjustment can be carried out using the R package called `MCPMod` available for download from the Comprehensive R Archive Network (CRAN) (Bornkamp, Pinheiro, and Bretz, 2014).

Two important observations can be made from these results. Firstly, unlike the variance optimality criteria, the complete and minimal power based optimality criteria are dependent on N via the z_i terms in (5.2) and (5.3). Secondly, unlike the variance optimality criteria, the complete and minimal power based optimality criteria are dependent on δ. The effect of this dependence on N and δ will be considered in turn in the remainder of this section and in the next section, respectively.

5.4.2 Constant heteroscedasticity and equal effects

In this section we consider the relationship between power optimal designs and variance optimal designs. The relationships are presented as conjectured approximations for large N based on numerical investigations discussed below.

Let $\gamma_0^{(C)}(N)$ be the control group proportion that maximises the complete power $P_C(N)$ in equation (5.2), for given N. Likewise, let $\gamma_i^{(C)}(N)$, $i=1,\ldots,k$ be the corresponding experimental group proportions. In the context of constant heteroscedasticity and equal effect sizes, numerical investigations suggest the following conjecture concerning the relationship between the complete power design and A-optimality.

Conjecture 5.4.1 *For the model with constant heteroscedasticity, and a constant treatment difference between the control and the comparator treatments, the control and comparator group proportions for the complete power design satisfy*

$$\gamma_0^{(C)}(N) \approx \frac{1}{1+r\sqrt{k}} = \gamma_0^{(A)} \qquad \textit{and} \qquad \gamma_i^{(C)}(N) \approx \frac{r}{\sqrt{k}(1+r\sqrt{k})} = \gamma_i^{(A)} \qquad \textit{for large } N.$$

Using the notations $\gamma_0^{(M)}(N)$ and $\gamma_i^{(M)}(N)$ to represent respectively, the control group and

experimental group proportions, $i = 1,\ldots,k$, that maximise the minimal power $P_M(N)$ in equation (5.3), numerical investigations also lead to an analogous conjecture for minimal power and E-optimality.

Conjecture 5.4.2 *For the model with constant heteroscedasticity, and a constant treatment difference between the control and the comparator treatments, the control group proportion for the minimal power design satisfies*

$$\gamma_0^{(M)}(N) \approx \frac{1}{1+r} = \gamma_0^{(E)} \qquad \textit{and} \qquad \gamma_i^{(M)}(N) \approx \frac{r}{k(1+r)} = \gamma_i^{(E)} \qquad \textit{for large } N.$$

5.4.3 Numerical comparisons

Graphical summaries supporting these conjectures are provided in Figure 5.4 for a $k = 3$ case (4 groups) and the corresponding numerical summaries are also provided in table format in the Web Appendix. The results suggest that all methods allocate the largest proportion to the control group and distribute the remaining experimental units equally between the comparator arms. As proposed in Conjectures 5.4.1 and 5.4.2, the numerical results demonstrate that the complete power design allocation ratios can be approximated by the A-optimal allocation ratios, and that the minimal power design allocation ratios can be approximated by the E-optimal allocation ratios.

5.5 Weighted optimality

A major difference between the variance optimal designs and the power optimal designs is that the magnitude of the treatment differences between the control and comparator arms which play a critical role in the determination of the allocation ratios under a power optimal design do not contribute to the determination of the allocation ratios under a variance optimal design. An approach to addressing this difference would be to consider a weighted version of the variance optimal designs, where a weighting is included in the variance-covariance, leading to the minimisation of $\Sigma_W = W^T \Sigma W$ for some $k \times 1$ weight matrix $W = (w_1,\ldots,w_k)^T$

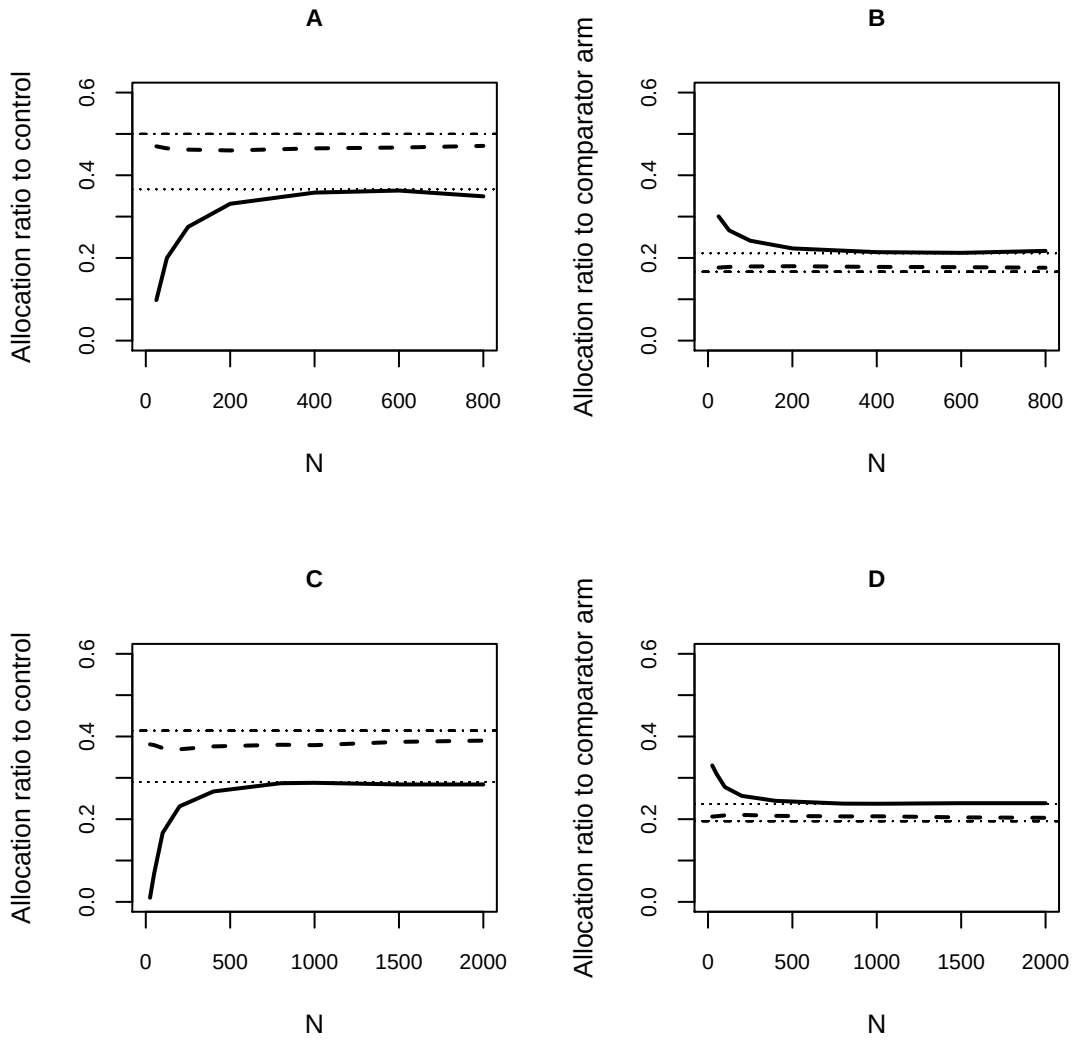

Figure 5.4: Optimal allocation proportions for the 4-group case under homoscedasticity (Panels A and B) and constant heteroscedasticity with $r_i = \sqrt{2}$ (Panels C and D). Allocation ratios to the control and comparator arms under the complete (solid line) and minimal power designs (dashed line), as functions of N, and A- (dotted line) and E-optimal (dot-dash line) designs are presented.

(Atkinson and Donev, 1992). Here,

$$\Sigma_W = \text{diag}\left(\frac{w_1^2 v_1}{\gamma_1}, \ldots, \frac{w_k^2 v_k}{\gamma_k}\right) + \frac{v_0}{\gamma_0} W W^{\mathrm{T}}, \tag{5.4}$$

which reduces to Σ as defined in (5.1) when $w_i = 1$, for $i = 1, \ldots, k$.

In the remainder of this section we discuss the weighted variant of the unified form discussed in Section 5.3.1 and propose conjectures relating to relationships between power optimal allocation ratios and the weighted variance optimal allocations.

5.5.1 Unified form

Morgan and Wang (2010) have previously explored the use of weighted variants of the *D*-, *A*-, and *E*-optimality criteria to assign differential interest in different comparisons including the general case which considers all pair-wise comparisons. Here we focus on the special case in which pair-wise comparisons with a common control are of interest. In this special case we are able to provide a unified form for the weighted variance optimal designs. We will henceforth refer to these weighted variants as the D_W-, A_W- and E_W-optimal designs. Using Σ_W, optimal allocation proportions for the D_W-, A_W- and E_W-optimal methods can be derived.

The generalisation of Proposition 5.3.1 for the weighted case can be stated as follows.

Proposition 5.5.1 *For the D_W-, A_W- and E_W-optimality criteria, the optimal design for the model with general heteroscedasticity is*

$$\gamma_{0W} = \frac{1}{1 + \overline{R}_W} \qquad \textit{and} \qquad \gamma_{iW} = \gamma_{0W} \hat{u}_i R(x_i, y_i, z_i) \qquad i = 1, \ldots, k$$

where $R(x,y,z) = [x + y(z-x)]^{-1}$ and $(\hat{u}_i, x_i, y_i, z_i)$ depends on the chosen optimality criterion according to Table 5.2, with $\overline{R}_W$ being the weighted average

$$\overline{R}_W = \sum_{i=1}^{k} \hat{u}_i R(x_i, y_i, z_i).$$

The proof of Proposition 5.5.1 is provided in the Web Appendix.

Two observations can be made based on a comparison of Propositions 5.3.1 and 5.5.1. Firstly, u_i and $\hat{u}_i$ specified in Table 5.2 are identical for the D-optimal design. That is, the D_W-optimal design is independent of W and is identical to the unweighted D-optimal design. This is consistent with the findings of Morgan and Wang (2010), and would generally imply that D-optimality is inappropriate for clinical trials when the design effect sizes are different across contrasts. A second observation is that under homoscedasticity and constant heteroscedasticity, the allocation ratio for the control group under the E_W-optimal design is independent of W. That is, it is the same as the allocation ratio under the unweighted E-optimal design. However, the allocation ratios to the comparator arms will not be the same, so E_W-optimality can be used to weight according to effect size. We now consider the relationship between weighted variance optimal designs and power optimal designs.

5.5.2 Constant heteroscedasticity and unequal effects

We have already seen that there seems to be a correspondence between power optimality and variance optimality for equal effect sizes. When effect sizes are unequal this correspondence is again maintained for the complete power design and A_W-optimality. As we will see, such correspondence for complete power is dependent on a particular choice of weights, namely, weighting inversely proportional to the squared effect size. With regards to minimal power, the relationship with E_W-optimality does persist in the control arm, but no longer holds in the comparator arms. The lack of correspondence in the non-control arms is a result of the difference in the underlying objective of the minimal power design and variance optimality. In the case of minimal power design, more resources are allocated to the comparator group which has the largest treatment difference compared with the control group as this maximises the power, or equivalently the probability of rejecting *at least one* of the H_{0i}. This is in contrast to the E_W-optimal design which tries to allocate more resources to the group with the smallest treatment difference compared with the control group, as it is this group that gets the smallest weight, and thereby will have a larger variance unless compensated

with more patients. Thus, in the case of constant heteroscedasticity and unequal effects, numerical investigations discussed below lead to the following two conjectures.

Conjecture 5.5.2 *For the model with constant heteroscedasticity, the control and comparator group proportions for the complete power design satisfies*

$$\gamma_0^{(C)}(N) \approx \gamma_0^{(A_W)} \qquad \textit{and} \qquad \gamma_i^{(C)}(N) \approx \gamma_i^{(A_W)} \qquad \textit{for large } N$$

when $W = (1/\delta_1^2, \ldots, 1/\delta_k^2)$.

Conjecture 5.5.3 *For the model with constant heteroscedasticity, the control group proportion for the minimal power design satisfies*

$$\gamma_0^{(M)}(N) \approx \frac{1}{1+r} = \gamma_0^{(E_W)} \qquad \textit{for large } N$$

for any general collection of weights W.

5.5.3 Numerical comparisons

Table 5.3 presents allocation ratios for the unweighted and weighted variance optimal designs with weights $W = (1/\delta_1^2, 1/\delta_k^2)$ together with power optimal allocation ratios in the scenario where $k = 2$ (3 groups). In the lower half of Table 5.3 we have chosen an extreme scenario of $\delta_2/\delta_1 = 4$, but Figure 5.5 and the examples presented in Section 5.6 are more moderate. These numerical results demonstrate that the approximations alluded to in the conjectures presented earlier hold under an assumption of homoscedasticity or constant heteroscedasticity. In particular the allocation ratios to the optimal complete power design are similar to the A-optimal allocation ratios when the treatment differences are the same across the treatment arms and to the A_W-optimal allocation ratios when the treatment differences are different across the treatment arms. Similarly, the E-optimal allocation ratios can be used to approximate the allocation ratios for a minimal power design when the treatment differences are the same across the treatment arms, but applies only to the control group

Table 5.3: Optimal allocation proportions for a study with $k = 2$ comparator arms with $r_i = 2$ under constant heteroscedasticity. Results for power optimal designs are presented for large N. The minimal power design presents allocation proportions without a correction for multiplicity of testing. The upper half of the table presents results when the treatment difference is the same between control and comparators ($\delta_2/\delta_1 = 1$) and the lower half presents the results when the difference between the control and the second comparator is 4 times larger than the treatment difference between the control and the first comparator ($\delta_2/\delta_1 = 4$).

Optimality criterion	Homoscedasticity			Constant heteroscedasticity		
	γ_0	γ_1	γ_2	γ_0	γ_1	γ_2
D	0.333	0.333	0.333	0.229	0.386	0.386
A	0.414	0.293	0.293	0.261	0.369	0.369
E	0.500	0.250	0.250	0.333	0.333	0.333
P_C	0.414	0.293	0.293	0.261	0.370	0.370
P_M	0.482	0.259	0.259	0.317	0.342	0.342
D_W	0.333	0.333	0.333	0.229	0.386	0.386
A_W	0.485	0.484	0.030	0.320	0.640	0.040
E_W	0.500	0.498	0.002	0.333	0.664	0.003
P_C	0.490	0.490	0.020	0.320	0.645	0.035
P_M	0.495	0.010	0.495	0.330	0.010	0.660

allocation ratios when the treatment differences are different across the treatment arms. As alluded to in Section 5.5.2, it can also be seen that the allocation across the comparator arms under E_W-optimality and minimal power design vary when the treatment differences are different across the treatment arms. Graphical presentations that illustrate the conjectures are presented in Figure 5.5 in the $k = 3$ case (4 group). The results in Table 5.3 also illustrate the independence of the allocation ratios under the D-optimal design to the weighting. This therefore limits the applicability of D-optimality in hypothesis testing situations such as clinical trials where the design treatment effect sizes are different, since power optimality will weight unequally when effect sizes are unequal.

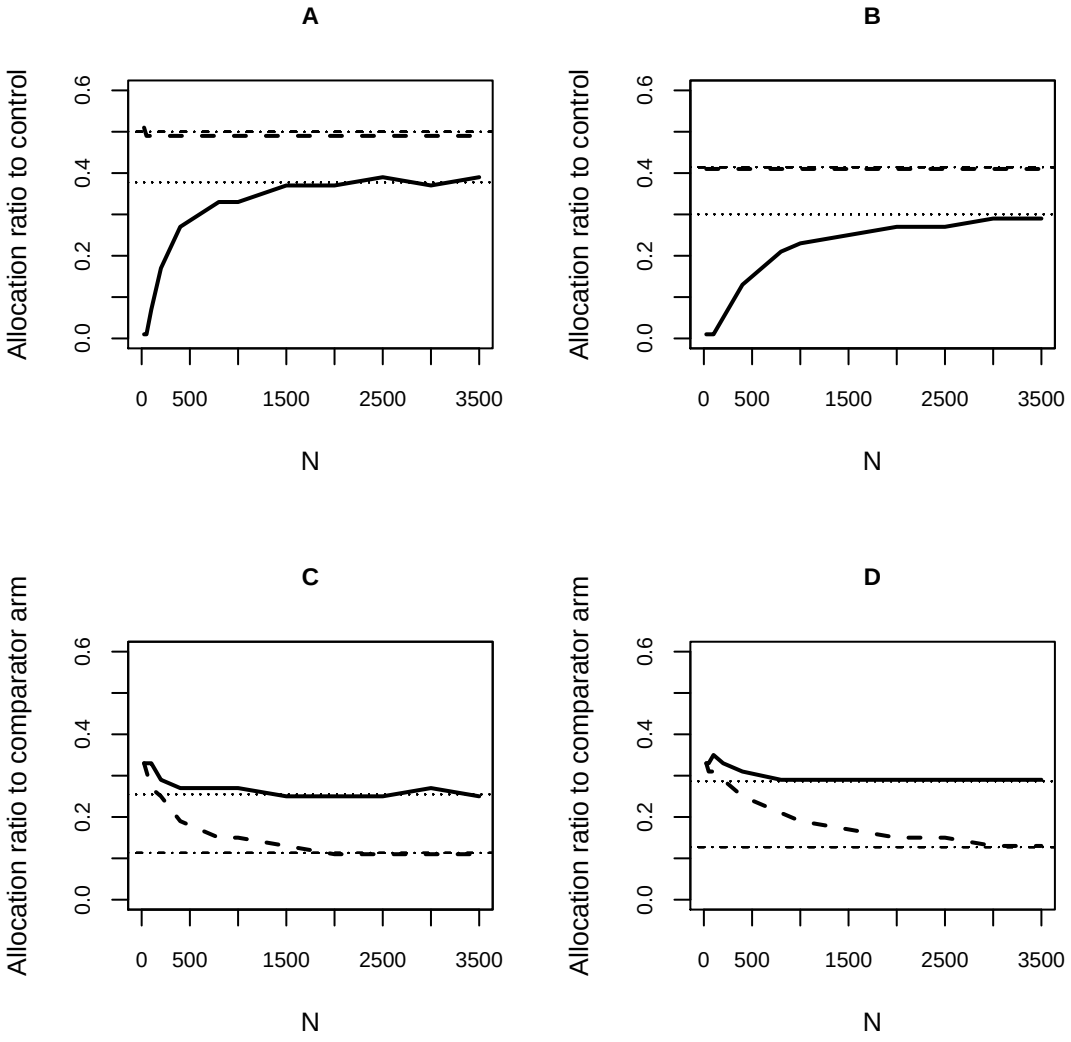

Figure 5.5: Optimal allocation proportions for the 4-group case under homoscedasticity and constant heteroscedasticity ($r_i = \sqrt{2}$) and unequal effect sizes $\delta = (0.2, 0.2, 0.3)$. All allocation ratios to power optimal designs are presented as functions of N. Panels A and B present the control group allocation ratios under complete power (solid line), minimal power (dashed line), A_W- (dotted line) and E_W-optimal (dot-dash line) designs with Panel A presenting these under homoscedasticity and Panel B under constant heteroscedasticity ($r_i = \sqrt{2}$). Panels C and D present the comparator arm allocation ratios with complete power and A_W-optimal design under homoscedasticity and constant heteroscedasticity ($r_i = \sqrt{2}$), respectively. In both these panels, the complete power design allocation proportion to the comparison with the smaller treatment difference ($\delta_1 = \delta_2 = 0.2$, solid line) and the larger treatment difference ($\delta_3 = 0.3$, dashed line) are presented together with the corresponding A_W-optimal design comparator group allocation ratios (the comparison $\delta_1 = \delta_2 = 0.2$ represented by the dotted line and the comparison $\delta_3 = 0.3$ represented by the dot-dash line).

5.6 Clinical trial examples

5.6.1 Example 1: NeoALTTO trial

The NeoALTTO trial was a multi-centre, open-label, phase 3 study that randomised patients with invasive breast cancer to trastuzumab (control arm), lapatinib, or lapatinib plus trastuzumab in a ratio of 1:1:1 (Baselga et al., 2012). The trial planned to enrol 450 patients to detect an increase of at least 17% in the pathological complete response (pCR) rate in the experimental groups compared with 25% in the trastuzumab group. This sample size was based on 80% power at a two-sided significance level of 2.5% for each of the two primary hypothesis, lapatinib plus trastuzumab versus trastuzumab, and lapatinib versus trastuzumab. This approach implies that the trial would have met its objective if at least one of these comparisons resulted in a statistically significant improvement in pCR. Therefore, this could have been framed as a family of hypotheses to be tested under a minimal power optimal design.

Table 5.4 presents the optimal allocation proportions to the control arm along with the power that may have been achieved if the hypotheses had been formulated under a complete power design or a minimal power design. Using the variance function as defined in Table 5.1 for a risk difference (17%), the variances can be calculated as $v_0 = 0.25 \times (1-0.25) = 0.1875$ and $v_1 = v_2 = 0.42 \times (1-0.42) = 0.2436$. This yields a constant heteroscedasticity design in which the standard deviation ratios are $r_1 = r_2 = 1.14$, with design treatment differences of 17% for each of the 2 hypotheses. The results in Table 5.4 suggest that the sample size of 450 patients under a balanced design is over-powered for a minimal power design, given the minimal power of 96.6% greatly exceeds the targeted 80%. The achieved complete power of 80.8% for the balanced design is consistent with the targeted 80%, however the use of complete power is not consistent with the way the primary comparisons were planned. Under a minimal power design, the NeoALTTO trial could have been designed with a sample size of 226 with 43.3% of the patients allocated to the control arm. If a Dunnett correction had been undertaken in the spirit of the Bonferroni adjustment for multiplicity that was done in the

NeoALTTO trial, a sample size of 284 patients, which is still smaller than the planned sample size in the NeoALTTO trial, would have sufficed. Finally, the results also demonstrate that using the allocation ratios of the A- and E-optimal designs achieves approximately the same power as the complete and minimal power designs, respectively, consistent with the relationships discussed in Section 5.4.2.

Table 5.4: Optimal allocation proportions using the design parameters of the NeoALTTO trial. The power optimal methods are based on a two-sided significance level of 5%. No correction for multiplicity has been applied to the minimal power design.

	N	γ_0	γ_1	γ_2	Complete power (%)	Minimal power (%)
Balanced	450	0.333	0.333	0.333	80.8	96.6
D	450	0.314	0.343	0.343	80.5	96.3
A	450	0.383	0.309	0.309	81.0	97.0
E	450	0.467	0.266	0.266	79.2	97.1
P_C	450	0.368	0.316	0.316	81.1	96.9
P_M	450	0.437	0.282	0.282	80.2	97.2
P_C	440	0.367	0.317	0.317	80.0	96.6
P_M	226	0.433	0.284	0.284	42.7	80.0

Hypothetically, a complete power design could have been used for the NeoALTTO trial if the combination therapy arm had been designated the common comparator and the primary objective had been to establish superiority of the combination over both mono-therapies. For this hypothetical scenario, Table 5.5 shows that a sample size of 417 patients (42.8% randomised to the combination therapy arm) would achieve a complete power of 80%. As an illustrative exercise, it is also possible to calculate p-values for this hypothetical trial using the observed rates of pCR from the NeoALTTO trial; 51.3% in the combination therapy arm, 24.7% in the lapatinib arm, and 29.5% in the trastuzumab arm. This results in a p-value of 0.0002 for the comparison between the combination therapy and trastuzumab and a p-value < 0.0001 for the comparison with lapatinib. Thus, the alternative design would have demonstrated superiority over both mono-therapies, but with a smaller sample size.

Table 5.5: Optimal allocation proportions using the combination therapy arm as the control (pCR rate=42%), the mono-therapy arms as the comparators (pCR rate=25%) and a two-sided significance level of 5%. No correction for multiplicity has been applied to the minimal power design.

	N	γ_0	γ_1	γ_2	Complete power (%)	Minimal power (%)
D	417	0.352	0.324	0.324	78.9	94.9
A	417	0.446	0.277	0.277	79.9	96.2
E	417	0.533	0.234	0.234	77.9	96.4
P_C	417	0.428	0.286	0.286	80.0	96.1
P_M	222	0.500	0.250	0.250	44.9	80.0

5.6.2 Example 2: Paliperidone palmitate in acutely exacerbated schizophrenia

The second example considers the design of a trial of 3 different doses of paliperidone palmitate (25, 100, and 150 mg) compared with placebo in adults with acutely exacerbated schizophrenia (Pandina et al., 2010). The primary endpoint was a mean change in the positive and negative syndrome scale (PANSS) from baseline to study end point. It was determined that 148 patients per arm (not adjusting for drop-out) provided 90% power to detect a difference of at least 9 points, assuming a standard deviation of 21 points, and a 2-sided overall significance level of 5%. The trial planned to use a Dunnett-Bonferroni-based parallel gatekeeping procedure to simultaneously adjust for multiplicity in the primary endpoint and a key secondary endpoint. We will use this example to evaluate how efficiency may have been increased with an unbalanced design, and in particular, how efficiency may have been increased if a dose-response relationship had been assumed, and demonstrate how Conjecture 5.5.2 holds in the latter case.

Table 5.6 presents the allocation ratios when the sample size used is large enough to ensure at least 90% power under complete power and minimal power designs and when various assumptions about the treatment differences relating to the dose-strength are made. Table 5.6 also presents the allocation ratios under the A_W- and E_W- designs for comparison with the complete and minimal power design allocations, respectively. The following observations

Table 5.6: Optimal allocation proportions using the design parameters of the paliperidone palmitate trial which assumed the same effect size across each comparison, and alternative designs which assume an increasing effect size with dose. The power optimal methods are based on a two-sided significance level of 5%. No correction for multiplicity has been applied to the minimal power design. Sample sizes that would have achieved at least 90% complete power or minimal power are presented.

	Design	N	γ_0	γ_1	γ_2	γ_3	Complete power (%)
$\delta = (9,9,9)$	Balanced	592	0.25	0.25	0.25	0.25	89.9
	A_W	564	0.37	0.21	0.21	0.21	90.0
	P_C	564	0.35	0.22	0.22	0.22	90.0
$\delta = (6,7.5,9)$	A_W	900	0.38	0.30	0.19	0.13	90.0
	P_C	900	0.37	0.31	0.19	0.13	90.0
$\delta = (7.5,9,10.5)$	A_W	608	0.37	0.28	0.20	0.14	90.2
	P_C	608	0.35	0.29	0.21	0.15	90.2
$\delta = (9,10.5,12)$	A_W	440	0.37	0.27	0.20	0.15	90.3
	P_C	440	0.37	0.27	0.21	0.15	90.3
	Design	N	γ_0	γ_1	γ_2	γ_3	Minimal power (%)
$\delta = (9,9,9)$	Balanced	592	0.25	0.25	0.25	0.25	99.6
	E_W	224	0.50	0.17	0.17	0.17	90.1
	P_M	224	0.44	0.19	0.19	0.19	90.4
$\delta = (6,7.5,9)$	E_W	228	0.50	0.31	0.13	0.06	73.2
	P_M	228	0.49	0.01	0.01	0.49	90.3
$\delta = (7.5,9,10.5)$	E_W	168	0.50	0.29	0.14	0.07	77.3
	P_M	168	0.49	0.01	0.01	0.49	90.5
$\delta = (9,10.5,12)$	E_W	128	0.50	0.27	0.14	0.09	79.7
	P_M	128	0.49	0.01	0.01	0.49	90.4

can be made from these results. The trial could have been designed under a complete power design with 564 patients or with 224 patients under a minimal power design (312 if Dunnett correction were applied) instead of 592 patients as was done in the trial. The allocation ratios under the complete power design and the A_W- design are similar across a range of treatment difference assumptions demonstrating how Conjecture 5.5.2 holds numerically. Furthermore, it is evident from Table 5.6 that consistently more patients are allocated to the treatment arm with the smallest expected treatment difference and the fewest patients to the arm with the largest expected treatment difference under the complete power and A_W-optimal designs. It can also be observed that while there is similarity between the allocation ratios to the control arm under minimal power and E_W- optimal designs, the allocation ratios to the non-control arms are different when the design treatment differences are assumed to

be different across the treatment groups. Indeed, the allocation ratios to the comparator arms suggest that a minimal power design is not sensible when the design treatment effect sizes are anticipated to be different across the contrasts, because it impractically concentrates the sample size in the one contrast that is most likely to achieve statistical significance.

5.7 Discussion

Unbalanced allocation of patients in single-control multiple-comparator clinical trials can result in a more efficient study design. In this paper we have provided a unification of optimal designs based on the D-, A- and E-optimality criteria, which facilitated comparisons between the three criteria. This has allowed us to identify systematic and potentially large differences between optimal designs based on the different criteria. These results have been presented in the context of a general model allowing for heteroscedasticity and a range of different treatment effect measures that include both continuous and binary outcomes.

Given the sensitivity of the optimal design to the chosen optimality criterion, we have argued that it is essential to match the optimality criterion to the objectives of the study. In most clinical trials, hypothesis testing is the primary objective and so variance optimality is less relevant than power optimality. We therefore studied optimisation of two versions of power, minimal and complete power, corresponding respectively to the situations where *any* or *all* of the k hypotheses must be rejected.

Although we found this approach is feasible in principle, we also found that the power optimal designs can be computationally intensive to determine. We therefore considered weighted versions of the variance optimality criteria which, like power optimal designs, allow the treatment difference to play a role in determining allocation ratios. We found that, at least for large sample sizes, the power optimal designs can be approximated by an appropriately chosen variance optimal design. In particular, the optimal complete power design is approximated by the A-optimal design under constant heteroscedasticity, with weighting by the inverse squared effect sizes. Furthermore, the optimal minimal power design is ap-

proximated by the E-optimal design assuming constant heteroscedasticity and equal effect sizes.

Based on our findings, we recommend the following rules of thumb in designing clinical trials involving multiple hypothesis tests with a common control. Firstly, if all of the k null hypotheses must be rejected for the study to be positive, then A-optimality should be used to determine the allocation proportions for each arm. This would involve using either the unweighted or weighted versions provided in Propositions 5.3.1 and 5.5.1, depending on whether or not the effect sizes are assumed to be equal for all comparisons. If they are unequal, then weighting by the inverse squared effect sizes should be used in Proposition 5.5.1 to retain a correspondence with optimising complete power.

Secondly, if the study would be positive if any of the k null hypotheses were rejected, then E-optimality should be used to determine the allocation proportions for each arm. This is straightforward for the case of equal effect sizes using Proposition 5.3.1, in which case there will be a correspondence with optimising minimal power. In the case of unequal effect sizes, optimising minimal power is not recommended since it leads to impractical designs that essentially omit treatment groups with smaller effect sizes. In this case there is no theoretical basis for any particular weights in the E_W-optimality criterion, although weighting by the inverse squared effect sizes produced sensible results in the two applications that we considered. For both of the scenarios discussed above, once the allocation proportions have been determined then the sample size N should be determined using the appropriate version of power in either (5.2) or (5.3).

Finally, we recommend that the D-optimality criterion should not be used in designing single-control multiple comparator trials that have hypothesis testing as the main objective. This is because D-optimality does not correspond to optimising either version of power, and it does not allow the allocation proportions to reflect differences in the treatment effects through weighting.

Conflict of Interest The authors have declared no conflict of interest.

5.8 References

Atkinson, A. C. and A. N. Donev (1992). *Optimum Experimental Designs*. Oxford, UK: Clarendon Press.

Baselga, J. et al. (2012). Lapatinib with trastuzumab for HER2-positive early breast cancer (NeoALTTO): a randomised, open-label, multicentre, phase 3 trial. *Lancet* **379**: 633–640.

Bornkamp, B., J. Pinheiro, and F. Bretz (2014). **MCPMod** : Design and analysis of dose-finding studies. R package version 1.0-8. http://CRAN.R-project.org/package=MCPMod.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* **50**: 1096–1121.

Fleiss, J. L. (1986). *The Design and Analysis of Clinical Experiments*. New York, USA: Wiley.

George, S. L. and M. M. Desu (1974). Planning the size and duration of a clinical trial studying the time to some critical event. *Journal of Chronic Diseases* **27**: 15–24.

Hedayat, A. S., M. Jacroux, and D. Majumdar (1988). Optimal designs for comparing test treatment with controls. *Statistical Science* **3**: 462–491.

Marschner, I. C. (2007). Optimal design of clinical trials comparing several treatments with a control. *Pharmaceutical Statistics* **6**: 23–33.

Morgan, J. P. and X. Wang (2010). Weighted optimality in designed experimentation. *Journal of the American Statistical Association* **105**: 1566–1580.

Pandina, G. J., J. P. Lindenmayer, J. Lull, P. Lim, S. Gopal, V. Herben, V. Kusumakar, E. Yuen, and J. Palumbo (2010). A randomized, placebo-controlled study to assess the efficacy and safety of 3 doses of paliperidone palmitate in adults with acutely exacerbated schizophrenia. *Journal of Clinical Psychopharmacology* **30**: 235–244.

Westfall, P. H., R. D. Tobias, D. Randall, D. Rom, R. D. Wolfinger, and Y. Hochberg (1999). *Multiple Comparisons and Multiple Tests*. Cary, USA: SAS Institute Inc.

Wong, W. K. and W. Zhu (2008). Optimum treatment allocation rules under a variance heterogeneity model. *Statistics in Medicine* **27**: 4581–4595.

Zhu, W. and W. K. Wong (2000). Optimal treatment allocation in comparative biomedical studies. *Statistics in Medicine* **19**: 639–648.

5.A Appendix: web-based supporting materials for "Design of clinical trials involving multiple hypothesis tests with a common control"

5.A.1 Proof for Proposition 5.3.1

This follows as a special case of the proof for Proposition 5.5.1 below.

5.A.2 Proof for Proposition 5.3.2

Assume constant heteroscedasticity, that is $r_i = r$ for $i = 1, \ldots, k$. Then the optimal proportions allocated to the control group under the D- and A-optimality criteria would be given by:

$$\begin{aligned} \gamma_0^{(A)} &= \left[1 + r\sqrt{k}\right]^{-1}, \qquad \text{and} \\ \gamma_0^{(D)} &= \left[1 + \frac{kr^2}{k\left(r^2-1\right)\gamma_0^{(D)} + 1}\right]^{-1} \end{aligned} \tag{5.A.1}$$

It can be seen that equation (5.A.1) is equivalent to a quadratic equation in $\gamma_0^{(D)}$ with two real roots

$$\frac{-(k+1) \pm \sqrt{4kr^2 + (k-1)^2}}{2k\left(r^2-1\right)}.$$

However, it can be shown the the root $\frac{-(k+1)-\sqrt{4kr^2+(k-1)^2}}{2k\left(r^2-1\right)}$ returns a value for $\gamma_0^{(D)}$ that is either < 0 or > 1, neither of which is the solution since $\gamma_0^{(D)}$ is a proportion, or is undefined when $r = 1$. Thus, the solution for $\gamma_0^{(D)}$ is $\frac{-(k+1)+\sqrt{4kr^2+(k-1)^2}}{2k\left(r^2-1\right)}$. Substituting this solution back into the right hand side of (5.A.1) gives

$$\gamma_0^{(D)} = \left[1 + r\sqrt{k}F_{rk}\right]^{-1}$$

where

$$F_{rk} = \frac{2r\sqrt{k}}{1-k+\sqrt{4kr^2+(k-1)^2}}. \tag{5.A.2}$$

The desired result therefore follows if we can show that $F_{rk} > 1$ since that would imply

$$\gamma_0^{(D)} < \left[1+r\sqrt{k}\right]^{-1} = \gamma_0^{(A)}.$$

The required inequality, $F_{rk} > 1$, follows from the binomial product rule which implies

$$\sqrt{4kr^2+(k-1)^2} < 2r\sqrt{k}+(k-1),$$

and hence the denominator of (5.A.2) is less than the numerator $2r\sqrt{k}$ which completes the proof.

5.A.3 Proof for Proposition 5.3.3

Let $\overline{R}^{(A)}$ and $\overline{R}^{(E)}$ be the $\overline{R}$ from Proposition 3.1 under the A- and E-optimality criteria, respectively.

We will first focus on the inequality $\gamma_0^{(A)} < \gamma_0^{(E)}$ by establishing that $\overline{R}^{(A)} > \overline{R}^{(E)}$ for $\max(r_1,\ldots,r_k) \leq 1$. The desired result will follow if it can be shown that $\overline{R}^{(A)} - \overline{R}^{(E)} > 0$ where

$$\begin{aligned}\overline{R}^{(A)} - \overline{R}^{(E)} &= \frac{\sum_{i=1}^k r_i}{\sqrt{k}} - \frac{\sum_{i=1}^k r_i(1+r_i)}{\sum_{i=1}^k (1+r_i)} \\ &= \frac{\left(\sum_{i=1}^k r_i\right)^2 + \sqrt{k}(\sqrt{k}-1)\sum_{i=1}^k r_i - \sqrt{k}\sum_{i=1}^k r_i^2}{\sqrt{k}\left(k+\sum_{i=1}^k r_i\right)}.\end{aligned} \tag{5.A.3}$$

Since the denominator of (5.A.3) is positive, we need to show that the numerator is positive.

$$
\begin{aligned}
& \left(\sum_{i=1}^{k} r_i\right)^2 + \sqrt{k}(\sqrt{k}-1)\sum_{i=1}^{k} r_i - \sqrt{k}\sum_{i=1}^{k} r_i^2 \\
> \;& (1-\sqrt{k})\sum_{i=1}^{k} r_i^2 + \sqrt{k}(\sqrt{k}-1)\sum_{i=1}^{k} r_i \\
= \;& \sum_{i=1}^{k}\left(r_i\sqrt{k}(\sqrt{k}-1)\right) - (\sqrt{k}-1)\sum_{i=1}^{k} r_i^2 \\
> \;& \sum_{i=1}^{k} r_i^2\left(k-\sqrt{k}-\sqrt{k}+1\right) \qquad \text{as } r \in (0,1] \\
= \;& \sum_{i=1}^{k} r_i^2(\sqrt{k}-1)^2 \\
> \;& 0
\end{aligned}
$$

That is, $\overline{R}^{(A)} > \overline{R}^{(E)}$ for $\max(r_1,\ldots,r_k) \leq 1$ and consequently, $\gamma_0^{(A)} < \gamma_0^{(E)}$.

We now turn our attention to establishing the inequality $\gamma_0^{(D)} < \gamma_0^{(A)}$. First define the function f such that $\gamma_0^{(D)}$ is the solution of the equation

$$
\gamma_0^{(D)} = f\left(\gamma_0^{(D)}\right) = \left[1 + \frac{kr^2}{k\left(r^2-1\right)\gamma_0^{(D)} + 1}\right]^{-1}.
$$

Then f is a decreasing function of $\gamma_0^{(D)}$ when $r_i < 1$. Therefore $\gamma_0^{(D)} < \frac{1}{k}$ since $f\left(\frac{1}{k}\right) = \frac{1}{1+k} < \frac{1}{k}$. Furthermore, both $\gamma_0^{(D)}$ and $\gamma_0^{(A)}$ are decreasing functions of r_i. Thus for $r_i \leq 1$, $\frac{1}{k} \geq \gamma_0^{(D)} \geq \frac{1}{1+k}$ and $\gamma_0^{(A)} \geq \frac{1}{1+\sqrt{k}}$. It follows that $\gamma_0^{(D)} < \gamma_0^{(A)}$ when $k > 2$ as $\frac{1}{k} < \frac{1}{1+\sqrt{k}}$ when $k > 2$. Showing that this also holds for $k = 2$ requires further proof as $\frac{1}{3} \leq \gamma_0^{(D)} < \frac{1}{2}$ and $\gamma_0^{(A)} \geq \frac{1}{1+\sqrt{2}}$ so the interval $\left[\frac{1}{1+\sqrt{2}}, \frac{1}{2}\right)$ can contain both $\gamma_0^{(D)}$ and $\gamma_0^{(A)}$. Let $R_m = \min\{r_1, r_2\}$. Since $\gamma_0^{(D)}$ is a decreasing function of r_i, representing $\gamma_0^{(D)}$ as a function of r_1 and r_2 leads to

$$
\gamma_0^{(D)} = \gamma_0^{(D)}(r_1, r_2) \leq \gamma_0^{(D)}(R_m, R_m) = \gamma_{0m}^{(D)}. \tag{5.A.4}
$$

Next we will show that $\gamma_{0m}^{(D)} \leq \gamma_0^{(A)}$ which then implies $\gamma_0^{(D)} \leq \gamma_0^{(A)}$ using (5.*A*.4). Firstly,

note that by inverting $\gamma_{0m}^{(D)} = \gamma_0^{(D)}(R_m, R_m)$ we have

$$R_m = \frac{\sqrt{(2\gamma_{0m}^{(D)} - 1)(\gamma_{0m}^{(D)} - 1)}}{\gamma_{0m}^{(D)}\sqrt{2}}. \tag{5.A.5}$$

Next let $R_d = |r_2 - r_1|$ so that

$$\gamma_0^{(A)} = \frac{1}{1 + \frac{R_m}{\sqrt{2}} + \frac{R_m + R_d}{\sqrt{2}}}. \tag{5.A.6}$$

Substituting R_m from equation (5.*A*.5) into equation (5.*A*.6) gives

$$\frac{\gamma_{0m}^{(D)}}{\gamma_0^{(A)}} = \gamma_{0m}^{(D)}\left(1 + \frac{R_d}{\sqrt{2}}\right) + \sqrt{(2\gamma_{0m}^{(D)} - 1)(\gamma_{0m}^{(D)} - 1)}. \tag{5.A.7}$$

Using the fact that $\gamma_{0m}^{(D)}$ is in the interval $[\frac{1}{3}, \frac{1}{2})$ and R_d is in the interval $[0, 1 - R_m]$ when $r_i \leq 1$ it follows that the right hand side of (5.*A*.7) is less than 1. That is, $\gamma_{0m}^{(D)} \leq \gamma_0^{(A)}$ which completes the proof since $\gamma_0^{(D)} \leq \gamma_{0m}^{(D)}$.

5.A.4 Proof for Proposition 5.3.4

Using the unified form of the optimal designs which was presented in Proposition 5.3.1, we have that

$$\begin{aligned}
\gamma_i^{(A)} &= \gamma_0^{(A)} r_i/\sqrt{k} \qquad \text{and} \\
\gamma_j^{(A)} &= \gamma_0^{(A)} r_j/\sqrt{k} \qquad \text{leading to} \\
g_{ij}^{(A)} &= r_i/r_j.
\end{aligned}$$

For $r_i/r_j \geq 1$, we have that $g_{ij}^{(A)} \geq 1$.

Similarly,

$$\begin{aligned} \gamma_i^{(E)} &= \gamma_0^{(E)} \frac{r_i(1+r_i)}{\sum_{l=1}^{k}(1+r_l)} \qquad \text{and} \\ \gamma_j^{(E)} &= \gamma_0^{(E)} \frac{r_j(1+r_j)}{\sum_{l=1}^{k}(1+r_l)} \qquad \text{leading to} \\ g_{ij}^{(E)} &= r_i(1+r_i)/r_j(1+r_j). \end{aligned}$$

For $r_i/r_j \geq 1$, it follows directly that $g_{ij}^{(E)} \geq g_{ij}^{(A)}$, thereby confirming the inequality $1 \leq g_{ij}^{(A)} \leq g_{ij}^{(E)}$.

5.A.5 Proof for Proposition 5.5.1

This section presents the proofs for the unified form for the weighted D_W-, A_W- and E_W-optimal designs. For simplicity of notation, we will use γ_i, for $i = 0, 1, \ldots, k$, in place of γ_{iW} throughout this section.

D-optimality

The determinant of Σ_W in equation (5.4) of the main paper is determined using the matrix determinant lemma presented by Harville (1997). This yields

$$\det(\Sigma_W) = \left[1 + \frac{v_0}{\gamma_0}\sum_{j=1}^{k}\frac{\gamma_j}{v_j}\right]\prod_{j=1}^{k}\frac{w_j^2 v_j}{\gamma_j}.$$

Let $f(\gamma) = \log[\det(\Sigma_W)]$. Applying the Lagrange multiplier approach with the constraint $g(\gamma) = \sum_{j=0}^{k}\gamma_j = 1$, we maximise the function $L(\gamma,\lambda) = f(\gamma) + \lambda\,(g(\gamma) - 1))$ which in the current context is given by

$$L(\gamma,\lambda) = \sum_{j=1}^{k}\log(w_j^2 v_j) - \sum_{j=1}^{k}\log(\gamma_j) + \log\left[1 + \frac{v_0}{\gamma_0}\sum_{j=1}^{k}\frac{\gamma_j}{v_j}\right] + \lambda\left\{\sum_{j=0}^{k}\gamma_j - 1\right\}. \quad (5.A.8)$$

Differentiating (5.A.8) with respect to γ_0 and γ_i and maximising these functions results in

two solutions for λ as follows:

$$\begin{aligned}
\frac{\partial L(\gamma,\lambda)}{\partial \gamma_i} &= \frac{-1}{\gamma_i} + \frac{v_0}{\gamma_0 v_i \left[1+\frac{v_0}{\gamma_0}\sum_{j=1}^{k}\frac{\gamma_j}{v_j}\right]} + \lambda = 0 \\
\lambda &= \frac{1}{\gamma_i} - \frac{v_0}{\gamma_0 v_i \left[1+\frac{v_0}{\gamma_0}\sum_{j=1}^{k}\frac{\gamma_j}{v_j}\right]} \\
&= \frac{1}{\gamma_i} - \frac{1}{r_i^2\left[\gamma_0+\sum_{j=1}^{k}\frac{\gamma_j}{r_j^2}\right]}, \qquad \text{and} \qquad (5.A.9) \\
\frac{\partial L(\gamma,\lambda)}{\partial \gamma_0} &= \frac{-v_0\sum_{j=1}^{k}\frac{\gamma_j}{v_j}}{\gamma_0^2\left[1+\frac{v_0}{\gamma_0}\sum_{j=1}^{k}\frac{\gamma_j}{v_j}\right]} + \lambda = 0 \\
\lambda &= \frac{v_0\sum_{j=1}^{k}\frac{\gamma_j}{v_j}}{\gamma_0^2\left[1+\frac{v_0}{\gamma_0}\sum_{j=1}^{k}\frac{\gamma_j}{v_j}\right]} \\
&= \frac{\sum_{j=1}^{k}\frac{\gamma_j}{r_j^2}}{\gamma_0\left[\gamma_0+\sum_{j=1}^{k}\frac{\gamma_j}{r_j^2}\right]}. \qquad (5.A.10)
\end{aligned}$$

Summing (5.A.9) over $j = 1,\ldots,k$ gives.

$$\begin{aligned}
\sum_{j=1}^{k}\lambda\gamma_j &= k - \frac{\sum_{j=1}^{k}\frac{\gamma_j}{r_j^2}}{\left[\gamma_0+\sum_{j=1}^{k}\frac{\gamma_j}{r_j^2\gamma_0}\right]} \\
\lambda(1-\gamma_0) &= k-\gamma_0\lambda \qquad \text{by substituting (5.A.10)} \\
\lambda &= k.
\end{aligned}$$

In order to arrive at an explicit solution for γ_i, the term $\sum_{j=1}^{k}\gamma_j/r_j^2$ needs to be determined. This is achieved by substituting $\lambda = k$ in (5.A.10) resulting in

$$\sum_{j=1}^{k}\frac{\gamma_j}{r_j^2} = \frac{k\gamma_0^2}{(1-k\gamma_0)}.$$

A solution for γ_i can now be determined by re-substitution of this result back into (5.A.9) which leads to

$$\gamma_i = \gamma_0\frac{1}{k}\left[\gamma_0+\frac{1}{r_i^2}\left(\frac{1}{k}-\gamma_0\right)\right]^{-1}.$$

Summing over the solution for γ_i from $i = 1, \ldots, k$ results in the following solution for γ_0

$$\gamma_0 = \frac{1}{\left\{1 + \sum_{i=1}^{k} \frac{1}{k}\left[\gamma_0 + \frac{1}{r_i^2}\left(\frac{1}{k} - \gamma_0\right)\right]^{-1}\right\}}.$$

This also demonstrates that the weights play no part in the determination of the allocation ratios under the *D*-optimal design.

A-optimality

The trace of the variance-covariance matrix Σ_W is given by

$$\sum_{j=1}^{k} \frac{w_j^2 v_0}{\gamma_0} + \sum_{j=1}^{k} \frac{w_j^2 v_j}{\gamma_j}.$$

As with the approach taken to determine γ_i for the *D*-optimal design, let $f(\gamma) = \text{trace}\,(\Sigma_W)$. Again, applying the Lagrange multiplier approach, we have that

$$L(\gamma, \lambda) = \sum_{j=1}^{k} \frac{w_j^2 v_0}{\gamma_0} + \sum_{j=1}^{k} \frac{w_j^2 v_j}{\gamma_j} + \lambda \left\{ \sum_{j=0}^{k} \gamma_j - 1 \right\}.$$

Differentiating this trace function with respect to γ_0 and γ_j and maximising these results in two solutions for λ as follows:

$$\begin{aligned}
\frac{\partial L(\gamma, \lambda)}{\partial \gamma_0} &= -\sum_{j=1}^{k} \frac{w_j^2 v_0}{\gamma_0^2} + \lambda = 0 \\
\lambda &= \sum_{j=1}^{k} \frac{w_j^2 v_0}{\gamma_0^2}, \qquad \text{and} \\
\frac{\partial L(\gamma, \lambda)}{\partial \gamma_i} &= -\frac{w_i^2 v_i}{\gamma_i^2} + \lambda = 0 \\
\lambda &= \frac{w_i^2 v_i}{\gamma_i^2}.
\end{aligned}$$

Setting these two solutions for λ equal to each other results in a solution for γ_i as follows

$$\gamma_i = \gamma_0 \frac{r_i w_i}{\sqrt{\sum_{j=1}^{k} w_j^2}}.$$

Using the constraint $\sum_{i=0}^{k} \gamma_i = 1$ results in the explicit solution

$$\gamma_0 = \frac{1}{\left[1 + \sum_{i=1}^{k} \frac{r_i w_i}{\sqrt{\sum_{j=1}^{k} w_j^2}}\right]}.$$

A special case of the weighted difference measure is one where all $w_i = 1$. In this special case we get the u_i represented by $1/k$ as presented in Table 5.1 of the main paper.

E-optimality

E-optimality stipulates the minimisation of the maximum eigenvalue of Σ_W which requires that there exists an eigenvalue τ that satisfies the condition $\det(\Sigma_W - \tau I) = 0$. Here I is the identity matrix. Again, using the matrix determinant lemma by Harville (1997) the determinant is given by

$$\det(\Sigma_W - \tau I) = \left[1 + \frac{v_0}{\gamma_0} \sum_{j=1}^{k} \frac{w_j^2 \gamma_j}{w_j^2 v_j - \tau \gamma_j}\right] \prod_{j=1}^{k} \left[\frac{w_j^2 v_j}{\gamma_j} - \tau\right]. \tag{5.A.11}$$

For ease of calculations, let s_i denote the information fraction n_i/v_i with the constraint $\sum_{j=0}^{k} \gamma_k = 1$ represented as $\sum_{j=0}^{k} s_j v_j = N$. It follows that the maximum eigen value of Σ_W is the solution over τ of the equation

$$\sum_{j=1}^{k} \frac{w_j^2 s_j}{\tau s_j / N - w_j^2} = s_0 \tag{5.A.12}$$

since the second term in (5.A.11) is the determinant of a diagonal matrix with non-zero elements (Rao, 1965). Letting $f(\gamma) = \sum_{j=1}^{k} \frac{w_j^2 s_j}{\tau s_j / N - w_j^2} - s_0$, and using the Lagrange multiplier

approach we have

$$L(s,\lambda,\tau) = \sum_{j=1}^{k} \frac{w_j^2 s_j}{\tau s_j/N - w_j^2} - s_0 + \lambda \left\{ \sum_{j=0}^{k} s_j v_j - N \right\}.$$

Differentiation of $L(s,\lambda,\tau)$ with respect to s_0 results in the simple solution, $\lambda = 1/v_0$. Substituting this value for λ into $\frac{\partial L(s,\lambda,\tau)}{\partial \gamma_i}$ results in the solution $s_i = \frac{w_i^2 N}{\tau}\left(1+\frac{1}{r_i}\right)$. A solution for s_i in terms of s_0 can be established if a solution for the nuisance parameter τ in terms of s_0 can be determined. This is achieved by substituting the result for s_i into (5.A.12) to find a solution of τ in terms of s_0 as follows

$$\tau = \frac{N}{s_0} \sum_{j=1}^{k} w_j^2 \left(1+r_j\right).$$

Substituting this result for τ leads to a solution for s_i in terms of s_0 as follows

$$\begin{aligned} s_i &= \frac{s_0 w_i^2 (1+r_i)}{r_i \sum_{j=1}^{k} w_j^2 \left(1+r_j\right)}, \qquad \text{or alternatively} \\ \gamma_i &= \gamma_0 r_i \frac{w_i^2 (1+r_i)}{\sum_{j=1}^{k} w_j^2 \left(1+r_j\right)}. \end{aligned}$$

The solution of γ_0 is now trivial and can be obtained from the constraint that $\sum_{j=1}^{k} \gamma_j = 1-\gamma_0$ as

$$\gamma_0 = \frac{1}{\left[1 + \sum_{i=1}^{k} r_i \frac{w_i^2(1+r_i)}{\sum_{j=1}^{k} w_j^2 (1+r_j)}\right]}.$$

A special case of the weighted difference measure is one where all $w_i = 1$. In this special case we get the u_i represented by $\frac{(1+r_i)}{\sum_{j=1}^{k}(1+r_j)}$ as presented in Table 5.1 of the main paper.

5.A.6 Additional numerical results

The numerical results relating to Figure 5.4 of the main paper is presented in this section.

Table 5.A.1: Optimal allocation proportions for a study with $k = 3$ comparator arms. Results for power optimal designs are presented for large N. The minimal power design presents allocation proportions without a correction for multiplicity of testing. Here $\delta = (0.5, 0.5, 0.5)$ and $r_i = \sqrt{2}$ under constant heteroscedasticity.

Optimality criterion	Homoscedasticity				Constant heteroscedasticity			
	γ_0	γ_1	γ_2	γ_3	γ_0	γ_1	γ_2	γ_3
D	0.250	0.250	0.250	0.250	0.215	0.262	0.262	0.262
A	0.366	0.211	0.211	0.211	0.290	0.237	0.237	0.237
E	0.500	0.167	0.167	0.167	0.414	0.195	0.195	0.195
P_C	0.367	0.211	0.211	0.211	0.285	0.238	0.238	0.238
P_M	0.476	0.175	0.175	0.175	0.391	0.203	0.203	0.203

5.A.7 Additional references

Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective*. New York, USA: Springer-Verlag.

Rao, C. R. (1965). *Linear Statistical Inference and its Application*. New York, USA: John Wiley & Sons.

Chapter 6

Conclusions

6.1 Summary of research

This thesis has presented statistical investigations of bias, inefficiency and misinterpretation associated with randomised clinical trials (RCTs). Specifically, three areas related to the design and analysis of RCTs have been studied:

1. bias in treatment effect estimates resulting from excluding early truncated trials from meta-analyses;

2. quantifying the expected chance variation in treatment differences between subgroups to avoid misinterpreting observed variation across subgroups, with a particular focus on country-specific analyses in multi-country RCTs; and

3. efficient design of single-control multiple-comparator trials through optimal unbalanced allocation.

This chapter presents a summary of the major findings and proposes some areas for future work that could be a natural follow on to the research presented in this thesis.

6.1.1 Meta-analysis and interim monitoring

The inclusion in meta-analysis of trials truncated early due to benefit has led to much debate with regards to the potential for overestimation of treatment effect estimates. This thesis has investigated the effect of excluding such studies from meta-analyses, which is an approach that has been advocated by some researchers (Bassler et al., 2013). The estimation

and information biases resulting from this approach were quantified theoretically in the simplest case where a study subjected to a single interim analysis continues to the final analysis without truncation, and through simulations in the cases with more than one interim analysis. Importantly, it was demonstrated that meta-analyses of non-truncated studies leads to underestimation of the treatment effect and overestimation of the statistical information. This has a problematic consequence in meta-analyses when estimates are weighted by the inverse-variance method, namely, that greater weighting is given to the most biased estimates. Indeed, it was observed in the simulation studies presented that the magnitude of the bias increased both when the studies were subjected to more frequent interim monitoring and when the proportion of the studies subjected to interim monitoring that were included in the meta-analyses increased. Although it might seem practicable to exclude from meta-analyses all studies subjected to interim monitoring, whether or not they terminated early, the resulting loss of efficiency in estimation can be considerable. This thesis concludes that the strategy that is most appropriate is to include all studies in evidence synthesis, both truncated and non-truncated. This is supported by the simulation studies presented which demonstrate that while excluding truncated studies led to substantial biases, the inclusion of all studies, both those that truncated early and those that did not, resulted in effectively unbiased estimates.

Since the publication of Schou and Marschner (2013), further discussion of estimation bias in truncated trials has appeared in the literature. Of particular note is the paper by Senn (2014) which cites the work presented in this thesis (Schou and Marschner, 2013) and explores issues that are directly related to the research presented in this thesis. While the theoretical component of Schou and Marschner (2013) focussed on quantifying the estimation and information biases in a trial subjected to a single interim analysis, Senn (2014) focuses on illustrating that the overall expectation of such a trial would indeed be unbiased as it is the sum of the expectation of the trial stopping early and that of the trial running its full course. Senn (2014) uses this point to demonstrate that unbiased estimates of treatment effect are achieved in information-weighted fixed-effect meta-analyses when trials subjected to a sin-

gle interim analysis, both truncated and non-truncated, are included. Senn (2014) further argues that this also applies in the case of multiple looks. The conclusions of Senn (2014) are therefore consistent with those of Schou and Marschner (2013), and provide further support for the conclusions reached in this thesis.

6.1.2 Subgroup analyses

Inadequate powering of tests of interaction is often a limitation encountered in subgroup analyses of clinical trials, and this can lead to interpretation difficulties when there is apparent heterogeneity between treatment effects across subgroups. While the focus of this thesis has been on country-specific subgroups that are often defined in multi-country RCTs, the methodology is readily applicable to other types of subgroup analyses. This thesis presents a suite of graphical tools which can provide a multi-faceted visual assessment of the extent to which the observed treatment effect differences align with those that would be expected under an assumption of treatment effect homogeneity. As it is likely that the limitation of low power for a test of interaction will also affect any new technique for assessing heterogeneity, this approach is entirely non-inferential. However, it does equip researchers, including non-statisticians, with a useful set of graphical presentations that are easy to understand and interpret. Furthermore, the utilisation of these tools at the design stage of a study can help benchmark expectations and pre-empt any over-interpretation at the analysis stage. As discussed in Chapter 4, these methods are accessible via the package **subgroup**, which has been included in the Comprehensive R Archive Network (CRAN) (Schou, 2014).

6.1.3 Optimal design

Single-control multiple-comparator trials offer both ethical and resource efficiencies by making effective therapies available sooner. However, the medical literature suggests that in the majority of cases, the design of such trials fails to harness efficiency gains that can be attained through unbalanced allocation of patients to the treatment groups.

Chapter 5 explored such unbalanced designs that could be achieved through variance opti-

mal designs and power optimal designs. In particular, three variance optimal designs based on the D-, A- and E-optimality criteria, with and without weighting were considered. These designs were unified under a single form such that the allocation ratios to each arm could be determined as a function of the number of comparator arms and the ratio of the standard deviation between each comparator arm and the common control. This unification allowed the sensitivity of the design to the chosen variance optimality criterion to be evaluated through a comparison of the way in which each method allocates the available resources to the control and comparator arms.

The results demonstrated that systematic orderings exist in the way the optimal designs allocate patients to the control and comparator arms for large sub-classes of models depending on the ratio of the standard deviation between the comparator and control arms of the trial. This dependency of the design allocation ratio on the chosen optimality criterion led to the proposal that power optimisation is more appropriate in the clinical trials setting where hypothesis testing is usually the focus. In this context, minimal and complete power optimal designs were discussed, and numerical examples were provided to support the use of approximate power optimal designs based on an appropriately chosen variance optimal design. These approximations are convenient as allocation ratios for exact power optimal designs are usually more complex to determine. These results led to some general guidelines on the design of single-control multiple-comparator trials. Specifically, it was noted that in a trial where all hypotheses had to be rejected for the overall study to be positive, that is, a complete power design, the A-optimal allocation ratios could be used as an approximation, with weighting by the inverse squared effect sizes. If rejection of any one of the null hypothesis would result in a positive study, that is, a minimal power design, the E-optimal allocation ratios could be used as an approximation. However, if the design treatment differences are unequal across the hypotheses, optimising minimal power is not sensible as it concentrates its effort on the hypothesis with the largest design treatment difference, and leads to the impractical scenario where treatment groups with the smaller effect sizes are essentially omitted. Finally, it was noted that the D-optimal design does not correspond to

either version of power optimality, and it ignores the design treatment differences even under weighting. For this reason, it was concluded that the *D*-optimality criterion is not suited for designing single-control multiple comparator trials that have hypothesis testing as the main objective.

6.2 Future research directions

This section discusses some potential areas for future work that could follow on from the research presented in this thesis.

6.2.1 Estimation biases due to interim monitoring

Since the publication of the research presented in Chapter 2, some further advances have been made that could form the basis of future research leading on from this thesis.

Chapter 2 discussed how estimates of treatment effect resulting from studies subjected to interim monitoring may be biased. In particular, it discussed how estimates from studies that stopped early overestimate the treatment effect, while studies that continue to the final analysis underestimate the treatment effect. The magnitude of these biases were theoretically formulated in the special case of a study with two analyses, one interim and one final, and validated through simulation studies. This section discusses some further advances that have been made since the publication of Schou and Marschner (2013). Specifically, it presents theoretical formulae for the estimation bias in trials subjected to more than one interim analysis. These theoretical formulae may be useful for the purpose of bias-adjusted estimation.

Several researchers have explored the issue of estimation bias in trials subjected to interim monitoring, and have proposed various bias-adjusted estimators. The first of these was Whitehead (1986), who suggested an iterative approach in which the bias of the maximum likelihood estimator (MLE) is subtracted from the observed value of the MLE. This approach was subsequently developed further from a computational and inferential perspec-

tive by Todd et al. (1996). Other approaches are based on existing methods for confidence interval estimation after a sequential study, such as the median unbiased estimate or mid-point estimate (Kim, 1989). Further approaches make use of the general concept of Rao-Blackwellization applied to the unbiased estimator obtained at the first interim analysis (Emerson and Fleming, 1990; Liu and Hall, 1999; Emerson and Kittelson, 1997). Importantly, all of these approaches are unconditional approaches in that they seek to adjust for the bias without conditioning on the stopping stage of the study. Fan et al. (2004) argued that a conditional approach to estimation may offer some advantages. This approach involves conditioning on the stopping stage and adjusting for the conditional bias given the stopping stage. It is this type of conditional adjustment that is considered further in this section.

Chapter 2 studied the estimation bias in a trial with a single interim analysis, using the conditional bias given the stopping stage. The quantification of the conditional bias in the more general case of a trial that is subjected to more than one interim analysis requires some further definitions. To this end, in the notation of Chapter 2, let $K+1$ denote the number of analyses, with K denoting the number of interim analyses. The information available at analysis k which was represented by $I^{(k)}$ in Chapter 2 is denoted by I_k in this chapter; that is, $I_k \equiv I^{(k)}$. Here we only consider a single study, and the subscript m used previously to denote the study number in a meta-analysis is therefore not applied. Recall that, as defined in Chapter 2, $\mathbf{D} = (D^{(1)}, \ldots, D^{(K)}, D^*)$ is the vector of treatment differences at the interim and final analyses, which has a $(K+1)$-dimensional multivariate normal distribution as defined in (2.7). For notational purposes we will let $D^{(K+1)} \equiv D^*$ in this section so $\mathbf{D} = (D^{(1)}, \ldots, D^{(K)}, D^{(K+1)})$. Consider now the multivariate distribution of the $(k-1)$-dimensional vector, $\mathbf{D}_k^{(i)} = (D^{(1)}, \ldots, D^{(i-1)}, D^{(i+1)} \ldots, D^{(k)})$, conditional on $D^{(i)}$. It can be shown that $\mathbf{D}_k^{(i)} \Big| D^{(i)}$ is multivariate normal

$$\mathbf{D}_k^{(i)} \Big| D^{(i)} \sim N\left(\boldsymbol{\delta}_k + [D^{(i)} - \delta] S_k^{(i)}, \Sigma_k^{(i)}\right) \qquad \text{where}$$

$$\Sigma_k^{(i)} = \begin{bmatrix} \frac{I_1-I_i}{I_1^2} & \frac{I_1-I_i}{I_1 I_2} & \cdots & \frac{I_1-I_i}{I_1 I_{i-1}} & \frac{I_1-I_i}{I_1 I_{i+1}} & \cdots & \frac{I_1-I_i}{I_1 I_k} \\ \frac{I_2-I_i}{I_1 I_2} & \frac{I_2-I_i}{I_2^2} & \cdots & \frac{I_2-I_i}{I_2 I_{i-1}} & \frac{I_2-I_i}{I_2 I_{i+1}} & \cdots & \frac{I_1-I_i}{I_2 I_k} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{I_k-I_i}{I_1 I_k} & \frac{I_k-I_i}{I_2 I_k} & \cdots & \frac{I_k-I_i}{I_{i-1} I_k} & \frac{I_k-I_i}{I_{i+1} I_k} & \cdots & \frac{I_k-I_i}{I_k^2} \end{bmatrix}.$$

Here $\boldsymbol{\delta}_k$ is a vector of length $k-1$ with all elements equal to δ and

$$S_k^{(i)} = \left(\frac{I_i}{I_1}, \ldots, \frac{I_i}{I_{i-1}}, \frac{I_i}{I_{i+1}}, \ldots, \frac{I_i}{I_k} \right).$$

The formulation of the required conditional expectation involves the evaluation of various density and cumulative distribution functions at the stopping boundaries. To this end, let the stopping boundary at the i^{th} analysis be c_i, so that the study stops for benefit if $D^{(i)} > c_i$. Furthermore, let $\phi^{(i)}$ denote the univariate normal density function of $D^{(i)}$, and let Θ_k denote the multivariate normal cumulative distribution function of $\mathbf{D}_k = (D^{(1)}, \ldots, D^{(k)})$. The multivariate distribution of $\mathbf{D}_k$ follows trivially from the multivariate distribution of $\mathbf{D}$ which was presented in (2.7). Finally, let $\Theta_k^{(i)}$ denote the multivariate normal cumulative distribution function of $\mathbf{D}_k^{(i)}$ as defined above. Using these definitions, the expectation of $D^{(k)}$ conditional on truncation at the k^{th} interim analysis can be given explicitly for $k = 1, \ldots, K$ by

$$\begin{aligned} E\left[D^{(k)} \middle| \text{truncation} \right] &= \delta \\ &- \frac{\sum_{i=1}^{k-1} \sqrt{I_i} \phi^{(i)}(c_i) \Theta_{k-1}^{(i)}(c_1, \ldots, c_{i-1}, c_{i+1}, \ldots, c_{k-1})}{I_k \left[\Theta_{k-1}(c_1, \ldots, c_{k-1}) - \Theta_k(c_1, \ldots, c_k) \right]} \\ &+ \frac{\sum_{i=1}^{k} \sqrt{I_i} \phi^{(i)}(c_i) \Theta_k^{(i)}(c_1, \ldots, c_{i-1}, c_{i+1}, \ldots, c_k)}{I_k \left[\Theta_{k-1}(c_1, \ldots, c_{k-1}) - \Theta_k(c_1, \ldots, c_k) \right]}. \end{aligned} \tag{6.1}$$

Likewise, the expectation of the treatment effect conditional on non-truncation is

$$E\left[D^{(K+1)}\middle|\text{non-truncation}\right] = \delta - \frac{\sum_{i=1}^{K}\sqrt{I_i}\phi^{(i)}(c_i)\Theta_K^{(i)}(c_1,\ldots,c_{i-1},c_{i+1},\ldots,c_K)}{I_{K+1}\Theta_K(c_1,\ldots,c_K)}. \quad (6.2)$$

It can be shown that in the special case of a trial with a single interim analysis, that is, $K = 1$, the relative biases derived from equations (6.1) and (6.2) correspond to equations (2.9) and (2.13) presented in Chapter 2.

In Table 6.1 a comparison of the theoretical and simulated biases is provided to check the validity of (6.1) and (6.2). These results show a close correspondence between the theoretical and simulated biases, which illustrates the validity of the theoretical expressions (6.1) and (6.2). Furthermore, they demonstrate that the underestimation bias in non-truncated studies increases steadily and can be in excess of 30% when $K > 3$, and that the smaller the information fraction at the time of analysis, the larger the overestimation bias in truncated studies. It is also apparent from this investigation that while the underestimation implied in Chapter 2 remains true in the general case for non-truncated RCTs, the estimation bias in truncated RCTs in more nuanced. Indeed, the overestimation in truncated RCTs wanes the later the look at which the study stops, and can even turn into an underestimation bias. This is in line with the observations made by Fan et al. (2004). For example, in Table 6.1 this is demonstrated in the case when $K+1 = 5$ and $k = 4$, as well as when $K+1 = 6$ and $k = 5$. Indeed, it can be shown that the underestimation bias is present already when $K+1 = 3$ if the interim analyses had taken place when the information fraction is large. For example, for a study monitored using the O'Brien Fleming boundaries at information fractions of $0.70, 0.85$, and 1, the underestimation at truncation when the information fraction is 0.85 is 12.1%. Therefore, the estimation bias in truncated studies can lead to either overestimation or underestimation depending on when the study stopped.

The theoretical expressions (6.1) and (6.2) are likely to be useful in bias-adjusted methods

Table 6.1: Treatment effect estimate and estimation bias expressed as percent overestimation or underestimation, conditional on truncation or non-truncation, of group sequential studies with K equally spaced interim analyses monitored using the O'Brien Fleming boundaries when the true treatment effect is $\delta = 0.25$ in trials with 90% power and one-sided significance level 2.5%.

Analyses	Interim look	Theoretical		Simulations		
$K+1$	k	Mean	% Bias	Samples	Mean	% Bias
2	1	0.373	49.2	3088	0.373	49.1
	2	0.222	-11.1	6912	0.223	-10.6
3	1	0.516	106.5	541	0.518	107.1
	2	0.305	21.9	5213	0.302	21.0
	3	0.191	-23.5	4246	0.191	-23.5
4	1	0.668	167.2	81	0.671	168.5
	2	0.374	49.7	2935	0.372	48.9
	3	0.265	5.9	4031	0.264	5.5
	4	0.174	-30.3	2953	0.175	-29.9
5	1	0.818	227.2	8	0.789	215.5
	2	0.446	78.4	1218	0.444	77.6
	3	0.315	26.0	3475	0.314	25.7
	4	0.241	-3.6	2829	0.242	-3.2
	5	0.163	-34.6	2470	0.163	-34.6
6	1	0.979	291.4	1	0.975	289.9
	2	0.521	108.6	702	0.521	108.5
	3	0.366	46.5	3563	0.367	46.7
	4	0.281	12.3	4307	0.281	12.3
	5	0.227	-9.3	3154	0.226	-9.5
	6	0.156	-37.4	3273	0.158	-36.8

for estimating treatment effects. In particular, Fan et al. (2004) made a compelling argument for modifying the unconditional approach of Whitehead (1986) and Todd et al. (1996) by using a conditional approach to bias adjustment. In particular their approach involves solving either of the following equations in δ

$$E\left[D^{(k)}\middle|\text{truncation}\right] = \hat{\delta} \qquad \text{or} \qquad E\left[D^{(K+1)}\middle|\text{non-truncation}\right] = \hat{\delta},$$

depending on whether the study was truncated or not. Here $\hat{\delta}$ is the crude MLE and the equations are solved for $\tilde{\delta}$, the conditional bias-adjusted estimate. However, Fan et al. (2004) solved these equations using numerical or simulation approximations to the theoretical bias.

Use of the exact theoretical expressions in (6.1) and (6.2) may offer various advantages including ease and efficiency of implementation. Preliminary numerical investigations indicate that this approach may hold some promise, but further research will be required to investigate this in more detail.

6.2.2 Quantifying heterogeneity of treatment effects

The methods presented in Chapter 3 offer a visual comparison of the observed variability in treatment effects with what would be expected under an assumption of treatment effect homogeneity across subgroups. However, in the event that the observed differences are more variable than the expected differences, a non-inferential measure of the extent to which these differences deviate from the assumption of homogeneity of treatment effect may be useful. A possible way to quantify this could be through the introduction of a random effects model. Recall that in Chapter 3 the treatment differences were defined as $D_r \sim N\left(\delta, s_r^2\right)$, $r = 1, \ldots, R$ in (3.1). Under a random effects model, these treatment differences can be defined as

$$\begin{aligned} D_r &\sim N\left(\Delta, s_r^2\right) \qquad \text{with} \\ \Delta &\sim N\left(\delta, \sigma_\delta^2\right) \qquad \text{implying} \\ D_r &\sim N\left(\delta, s_r^2 + \sigma_\delta^2\right) \qquad r = 1, \ldots, R. \end{aligned} \tag{6.3}$$

Under this parametrisation, σ_δ^2 would capture the extent of variation beyond what would be expected under homogeneity, and an estimate of σ_δ^2 could be used as a non-inferential measure to quantify the extent of this variation.

To illustrate how this might be used in practice, consider again the PLATO study which was discussed in Chapter 3. Chapter 3 concluded that overall, the variation in the country-specific treatment differences were consistent with what would be expected under an assumption of homogeneity of treatment effect. While this overall statement remains valid, it can be observed that the range of the observed treatment differences is slightly larger than the expected range of the treatment differences in the largest 10 countries of the PLATO

study as presented in Panel A of Figure 3.1. Here, the largest hazard ratio observed was 1.27 and the smallest was 0.59, a range of 0.767 on the log scale. The expected range, also on the log scale, was 0.691. That is, the observed range is slightly larger than the expected range under an assumption of homogeneity of treatment effect across the countries. Using the definition of the distribution of D_r presented in (6.3) it would be possible to estimate the value of σ_δ that would explain the increase in the observed range compared with the expected range. In this instance, letting the log of the hazard ratios in each country represent D_r, $r = 1, \ldots, 10$, the estimate of σ_δ is 0.109. That is, if $D_r \sim N\big(\log(\delta), s_r^2 + 0.109^2\big)$, the expected range and the observed range would be identical. This estimate of σ_δ could provide a non-inferential measure of the extent of departure from the assumption of homogeneity of country-specific treatment differences. Further exploration of the usefulness of such a measure, and its extension to incorporate information on all subgroup treatment effects rather than just the range, may be a worthwhile line of future research.

6.2.3 Further theory on single-control multiple-comparator trials

Numerical investigations were used in Chapter 5 to support the conjectures presented, which proposed that allocation ratios to the complete and minimal power optimal designs could be approximated by the A- and E-variance optimal design allocation ratios, respectively. Further theoretical investigations would be required to determine whether these conjectured approximations correspond to mathematical limits as the total sample size increases. If such relationships could be established, this would provide further theoretical basis for the guidelines provided in Chapter 5. Such theoretical investigations may therefore be a useful line of future research, although attempts thus far have been unsuccessful.

Chapter 5 also presented several propositions providing some general orderings of the allocation ratios for D-, A- and E-optimal designs. This discussion was limited to unweighted variance optimal designs. Therefore a natural follow-up would be to consider whether these propositions would hold in a more general context of weighted variance optimal designs, and if so, under what conditions on the weights these would hold. For example, under

constant heteroscedasticity, it is obvious that Proposition 5.3.2 would hold when the design treatment differences are equal across the hypotheses. However, these orderings may not hold when the design treatment differences are unequal. Thus, further investigations into potential orderings under weighted variance optimal designs may lead to useful insights.

Finally, it is noted there is scope to extend the work presented in Chapter 5 to accommodate studies subjected to sequential monitoring. Various authors have studied multi-stage studies in which some arms are dropped at interim analyses; see Stallard and Todd (2003), Kelly et al. (2005) and Stallard and Friede (2008). This type of RCT has recently been extended to allow for a power (or equivalently sample size) optimal design criterion by Wason and Jaki (2012). However, the computations involved are rather complex, requiring a stochastic search algorithm, and it may be that the variance optimality approximations discussed in Chapter 5 have applicability in that context as well.

6.3 Final remarks

Randomised clinical trials are a cornerstone of evidence-based medicine and public health. The RCT design is pivotal for both exploratory and confirmatory clinical research studies, as well as evidence synthesis through meta-analysis. Nonetheless, despite the widespread use and reporting of RCTs, there remain many methodological issues relevant to their design, analysis and interpretation. This thesis has carried out statistical investigations of three such areas.

The research presented here is based on theoretical and methodological investigations into the potential for bias, inefficiency and misinterpretation of RCTs. Some analysis and computational tools have also been presented, along with guidelines and recommendations for the valid design, analysis and interpretation of RCTs. Finally, some potential extensions and generalisations of the research have also been discussed. Together with the results presented here, these provide a framework for possible future investigations that go beyond the content of this thesis.

References

Arnold, B. C., N. Balakrishnan, and H. N. Nagaraja (2008). *A First Course in Order Statistics*. Philadelphia, USA: Society for Industrial and Applied Mathematics.

Atkinson, A. C. and A. N. Donev (1992). *Optimum Experimental Designs*. Oxford, UK: Clarendon Press.

Balakrishnan, N. (2007). Permanents, order statistics, outliers, and robustness. *Revista Matemática Complutense* **1**: 7–107.

Baselga, J. et al. (2012). Lapatinib with trastuzumab for HER2-positive early breast cancer (NeoALTTO): a randomised, open-label, multicentre, phase 3 trial. *Lancet* **379**: 633–640.

Bassler, D., M. Briel, V. M. Montori, M. Lane, P. Glasziou, Q. Zhou, D. Heels-Ansdell, S. D. Walter, and G. H. Guyatt (2010). Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. *Journal of the American Medical Association* **303**: 1180–1187.

Bassler, D., I. Ferreira-Gonzalez, M. Briel, D. J. Cook, P. J. Devereaux, D. Heels-Ansdell, H. Kirpalani, M. O. Meade, V. M. Montori, A. Rozenberg, H. J. Schünemann, and G. H. Guyatt (2007). Systematic reviewers neglect bias that results from trials stopped early for benefit. *Journal of Clinical Epidemiology* **60**: 869–873.

Bassler, D., V. M. Montori, M. Briel, P. Glasziou, S. D. Walter, T. Ramsay, and G. Guyatt (2013). Reflections on meta-analyses involving trials stopped early for benefit: Is there a problem and if so, what is it? *Statistical Methods in Medical Research* **22**: 159–168.

Bassler, D., V. M. Montori, M. Briel, P. Glasziou, Q. Zhou, and G. H. Guyatt (2008). Early stopping of randomized clinical trials for overt efficacy is problematic. *Journal of Clinical Epidemiology* **61**: 241–246.

Berry, S. M., B. P. Carlin, and J. Connor (2010). Bias and trials stopped early for benefit. *Journal of the American Medical Association* **304**: 156.

Bornkamp, B., J. Pinheiro, and F. Bretz (2014). **MCPMod** : Design and analysis of dose-finding studies. R package version 1.0-8. http://CRAN.R-project.org/package=MCPMod.

Buyse, M. and I. C. Marschner (2011). Assessment of statistical heterogeneity in the PLATO trial. *Cardiology* **118**: 138.

Chen, J., H. Quan, B. Binkowitz, S. P. Ouyang, Y. Tanaka, G. Li, S. Menjoge, and E. Ibia for the Consistency Workstream of the PhRMA MRCT Key Issue Team (2010). Assessing consistent treatment effect in a multi-regional clinical trial: a systematic review. *Pharmaceutical Statistics* **9**: 242–253.

Chen, J., H. Quan, P. Gallo, S. Menjoge, X. Luo, Y. Tanaka, G. Li, S. P. Ouyang, B. Binkowitz, E. Ibia, S. Talerico, and K. Ikeda (2011). Consistency of treatment effect across regions in multiregional clinical trials, part 1: design considerations. *Drug Information Journal* **45**: 595–602.

Chen, J., H. Zheng, H. Quan, G. Li, P. Gallo, S. P. Ouyang, B. Binkowitz, N. Ting, Y. Tanaka, X. Luo, and E. Ibia for the Society for Clinical Trials (SCT) Multi-Regional Clinical Trial Consistency Working Group (2013). Graphical assessment of consistency in treatment effect among countries in multi-regional clinical trials. *Clinical Trials* **10**: 842–851.

Cohen, A. C. (1991). *Truncated and Censored Samples: Theory and Applications*. New York, New York: Marcel Dekker.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* **50**: 1096–1121.

Durrett, R. (2010). *Probability: Theory and Examples*. Fourth. New York, New York: Cambridge University Press.

Emerson, S. S. and T. R. Fleming (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika* **77**: 875–892.

Emerson, S. S. and J. M. Kittelson (1997). A computationally simpler algorithm for the UMVUE of a normal mean following a group sequential trial. *Biometrics* **53**: 365–369.

Fan, X., D. L. DeMets, and K. K. G. Lan (2004). Conditional bias of point estimates following a group sequential test. *Journal of Biopharmaceutical Statistics* **14**: 505–530.

FDA (2010). Ticagrelor Secondary Review. http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/CardiovascularandRenalDrugsAdvisoryCommittee/UCM220192.pdf.

Fleiss, J. L. (1986). *The Design and Analysis of Clinical Experiments*. New York, USA: Wiley.

Freidlin, B. and E. L. Korn (2009). Stopping clinical trials early for benefit: impact on estimation. *Clinical Trials* **6**: 119–125.

Gallo, P., J. Chen, H. Quan, S. Menjoge, X. Luo, Y. Tanaka, G. Li, S. P. Ouyang, B. Binkowitz, E. Ibia, S. Talerico, and K. Ikeda (2011). Consistency of treatment effect across regions in multiregional clinical trials, part 2: monitoring, reporting and interpretation. *Drug Information Journal* **45**: 603–608.

George, S. L. and M. M. Desu (1974). Planning the size and duration of a clinical trial studying the time to some critical event. *Journal of Chronic Diseases* **27**: 15–24.

Goodman, S. N. (2008). Systematic reviews are not biased by results from trials stopped early for benefit. *Journal of Clinical Epidemiology* **61**: 95–96.

Green, S. J., T. R. Fleming, and S. Emerson (1987). Effects on overviews of early stopping rules for clinical trials. *Statistics in Medicine* **6**: 361–367.

Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective*. New York, USA: Springer-Verlag.

Hedayat, A. S., M. Jacroux, and D. Majumdar (1988). Optimal designs for comparing test treatment with controls. *Statistical Science* **3**: 462–491.

Hughes, M. D., L. S. Freedman, and S. J. Pocock (1992). The impact of stopping rules on heterogeneity of results in overviews of clinical trials. *Biometrics* **48**: 41–53.

Hughes, M. D. and S. J. Pocock (1988). Stopping rules and estimation problems in clinical trials. *Statistics in Medicine* **7**: 1231–1242.

Hung, H. M. J., S. J. Wang, and R. T. O'Neill (2010). Consideration of regional difference in design and analysis of multi-regional trials. *Pharmaceutical Statistics* **9**: 173–178.

Ibia, E. O. and B. Binkowitz (2011). Proceedings of the DIA Workshop on Multiregional Clinical Trials, October 26-27, 2010. *Drug Information Journal* **45**: 391–403.

ICH (1998). Statistical Principles for Clinical Trials E9. http://www.ich.org/products/guidelines/efficacy/article/efficacy-guidelines.html.

Jennison, C. and B. W. Turnbull (2000). *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, Florida: Chapman & Hall/CRC.

Kelly, P. J., N. Stallard, and S. Todd (2005). An adaptive group sequential design for phase II/III clinical trials that select a single treatment from several. *Journal of Biopharmaceutical Statistics* **15**: 641–658.

Kim, K. (1989). Point estimation following group sequential tests. *Biometrics* **45**: 613–617.

Li, Z., C. Chuang-Stein, and C. Hoseyni (2007). The probability of observing negative subgroup results when the treatment effect is positive and homogeneous across all subgroups. *Drug Information Journal* **41**: 47–56.

Liu, A. and W. J. Hall (1999). Unbiased estimation following a group sequential test. *Biometrika* **86**: 71–78.

Marschner, I. C. (2007). Optimal design of clinical trials comparing several treatments with a control. *Pharmaceutical Statistics* **6**: 23–33.

Marschner, I. C. (2010). Regional differences in multinational clinical trials: anticipating chance variation. *Clinical Trials* **7**: 147–156.

MERIT-HF Study Group (1999). Effect of metoprolol CR/XL in chronic heart failure: metoprolol CR/XL randomised intervention trial in congestive heart failure (MERIT-HF). *Lancet* **353**: 2001–2007.

Montori, V. M. et al. (2005). Randomized trials stopped early for benefit: a systematic review. *Journal of the American Medical Association* **294**: 2203–2209.

Morgan, J. P. and X. Wang (2010). Weighted optimality in designed experimentation. *Journal of the American Statistical Association* **105**: 1566–1580.

Muthén, B. (1990). Moments of the censored and truncated bivariate normal distribution. *British Journal of Mathematical and Statistical Psychology* **43**: 131–143.

Pandina, G. J., J. P. Lindenmayer, J. Lull, P. Lim, S. Gopal, V. Herben, V. Kusumakar, E. Yuen, and J. Palumbo (2010). A randomized, placebo-controlled study to assess the efficacy and safety of 3 doses of paliperidone palmitate in adults with acutely exacerbated schizophrenia. *Journal of Clinical Psychopharmacology* **30**: 235–244.

Pechtl, A. (1998). A note on the derivative of the normal distribution's logarithm. *Archiv der Mathematik* **70**: 83–88.

Pocock, S. J. and M. D. Hughes (1989). Practical problems in interim analyses with particular regard to estimation. *Controlled Clinical Trials* **10**: 209–221.

Quan, H., M. Li, J. Chen, P. Gallo, B. Binkowitz, E. Ibia, Y. Tanaka, S. P. Ouyang, X. Luo, G. Li, S. Menjoge, S. Talerico, and K. Ikeda (2010). Assessment of consistency of treatment effects in multiregional clinical trials. *Drug Information Journal* **44**: 617–632.

Quan, H., P. L. Zhao, J. Zhang, M. Roessner, and K. Aizawa (2010). Sample size considerations for Japanese patients in a multi-regional trial based on MHLW Guidance. *Pharmaceutical Statistics* **9**: 100–112.

R Development Core Team (2014). R: A Language and Environment for Statistical Computing. www.R-project.org. R Foundation for Statistical Computing: Vienna, Austria.

Rao, C. R. (1965). *Linear Statistical Inference and its Application*. New York, USA: John Wiley & Sons.

Rosenbaum, S. (1961). Moments of a truncated bivariate normal distribution. *Journal of the Royal Statistical Society* **23**.B: 405–408.

Rothwell, P. M. (2005). Subgroup analysis in randomised controlled trials: importance, indications and interpretation. *Lancet* **365**: 176–186.

Schou, I. M. (2014). subgroup: Methods for exploring treatment effect heterogeneity in subgroup analysis of clinical trials. R package version 1.1. http://CRAN.R-project.org/package=subgroup.

Schou, I. M. and I. C. Marschner (2011). Biases in clinical trials with sequential monitoring. *Trials* **12**.Suppl 1: A50.

Schou, I. M. and I. C. Marschner (2013). Meta-analysis of clinical trials with early stopping: an investigation of potential bias. *Statistics in Medicine* **32**: 4859–4874.

Schou, I. M. and I. C. Marschner (2015). Methods for exploring treatment effect heterogeneity in subgroup analysis: an application to global clinical trials. *Pharmaceutical Statistics* **14**: 44–55.

Senn, S. (2014). A note regarding meta-analysis of sequential trials with stopping for efficacy. *Pharmaceutical Statistics* **13**: 371–375.

Serebruany, V. L. (2010). Aspirin dose and ticagrelor benefit in PLATO: fact or fiction? *Cardiology* **117**: 280–283.

S+SeqTrial User's Manual (2002). Insightful Corporation. Seattle, Washington.

Stallard, N. and T. Friede (2008). A group-sequential design for clinical trials with treatment selection. *Statistics in Medicine* **27**: 6209–6227.

Stallard, N. and S. Todd (2003). Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine* **22**: 689–703.

Todd, S. (1997). Incorporation of sequential trials into a fixed effects meta-analysis. *Statistics in Medicine* **16**: 2915–2925.

Todd, S., J. Whitehead, and K. M. Facey (1996). Point and interval estimation following a sequential clinical trial. *Biometrika* **83**: 453–461.

Wallentin, L. et al. (2009). Ticagrelor versus clopidogrel in patients with acute coronary syndromes. *New England Journal of Medicine* **361**: 1045–1057.

Wang, R., S. W. Lagakos, J. H. Ware, D. J. Hunter, and J. M. Drazen (2007). Statistics in medicine - reporting of subgroup analyses in clinical trials. *The New England Journal of Medicine* **357**.21: 2189–2194.

Wason, J. M. S. and T. Jaki (2012). Optimal design of multi-arm multi-stage trials. *Statistics in Medicine* **31**: 4269–4279.

Wedel, H., D. DeMets, P. Deedwania, B. Fagerberg, S. Goldstein, S. Gottlieb, A. Hjalmarson, J. Kjekshus, F. Waagstein, and J. Wikstrand on behalf of the MERIT-HF Study Group (2001). Challenges of subgroup analyses in multinational clinical trials: experiences from the MERIT-HF trial. *American Heart Journal* **142**: 502–511.

Westfall, P. H., R. D. Tobias, D. Randall, D. Rom, R. D. Wolfinger, and Y. Hochberg (1999). *Multiple Comparisons and Multiple Tests*. Cary, USA: SAS Institute Inc.

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

Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials*. Second. Chichester, UK: Wiley.

Wilhelm, S. and B. G. Manjunath (2012). *tmvtnorm: Truncated multivariate normal and t distributions*. R package version 1.4-4.

Wittes, J. (2013). "Why is this subgroup different from all other subgroups? Thoughts on regional differences in randomized clinical trials". *Proceedings of the Fourth Seattle Symposium in Biostatistics: Clinical Trials*. Ed. by W. B. S. Fleming T. R. New York, USA: Springer: 95–115.

Wong, W. K. and W. Zhu (2008). Optimum treatment allocation rules under a variance heterogeneity model. *Statistics in Medicine* **27**: 4581–4595.

Zhu, W. and W. K. Wong (2000). Optimal treatment allocation in comparative biomedical studies. *Statistics in Medicine* **19**: 639–648.